A peer-reviewed version of this preprint was published in PeerJ on 4 May 2018.

Wangensteen OS, Palacín C, Guardiola M, Turon X. 2018. DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. PeerJ 6:e4705 https://doi.org/10.7717/peerj.4705

Metabarcoding littoral hard-bottom communities: unexpected diversity and database gaps revealed by two molecular markers

Owen S Wangensteen [1, 2] , **Creu Palacín** [3] , **Magdalena Guardiola** [1] , **Xavier Turon** Corresp. [1]

[1] Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB-CSIC), Blanes, Spain

[2] Ecosystems and Environment Research Centre, University of Salford, Salford, United Kingdom

[3] Department of Evolutionary Biology, Ecology and Environmental Sciences, and Biodiversity Research Institute (IRBio), University of Barcleona, Barcelona, Spain

Corresponding Author: Xavier Turon
Email address: xturon@ceab.csic.es

We developed a metabarcoding method for biodiversity characterization of structurally complex natural marine hard-bottom communities. Novel primer sets for two different molecular markers: the “Leray fragment” of mitochondrial cytochrome *c* oxidase, COI, and the V7 region of ribosomal RNA 18S were used to analyse eight different marine shallow benthic communities from two National Parks in Spain (one in the Atlantic Ocean and another in the Mediterranean Sea). Samples were sieved into three size fractions from where DNA was extracted separately. Bayesian clustering was used for delimiting molecular operational taxonomic units (MOTUs) and custom reference databases were constructed for taxonomic assignment. We found unexpectedly high values for MOTU richness, suggesting that these communities host a large amount of yet undescribed eukaryotic biodiversity. Significant gaps are still found in sequence reference databases, which currently prevent the complete taxonomic assignation of the detected sequences. Nevertheless, over 90% (in abundance) of the sequenced reads could be successfully assigned to phylum or lower taxonomical level. This identification rate might be significantly improved in the future, as reference databases are updated. Our results show that marine metabarcoding, currently applied mostly to plankton or sediments, can be adapted to structurally complex hard bottom samples, and emerges as a robust, fast, objective and affordable method for comprehensively characterizing the diversity of marine benthic communities dominated by macroscopic seaweeds and colonial or modular sessile metazoans, allowing for standardized biomonitoring of these ecologically important communities. The new universal primers for COI can potentially be used for biodiversity assessment with high taxonomic resolution in a wide array of marine, terrestrial or freshwater eukaryotic communities.

Metabarcoding littoral hard-bottom communities: unexpected diversity and database gaps revealed by two molecular markers

Owen S. Wangensteen[1,2], Creu Palacín[3], Magdalena Guardiola[1], and Xavier Turon[1]

[1] Center for Advanced Studies of Blanes (CEAB-CSIC), Girona, Spain

[2] Current address: Ecosystems and Environment Research Centre, University of Salford, UK

[3] Department of Animal Biology, University of Barcelona, Spain

Corresponding author: Xavier Turon

Email address: xturon@ceab.csic.es

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3429v1 | CC BY 4.0 Open Access | rec: 23 Nov 2017, publ: 23 Nov 2017

Keywords: biodiversity assessment, cytochrome *c* oxidase I, eukaryotic communities, marine benthic ecosystems, metabarcoding pipelines, ribosomal RNA 18S

Running title: Metabarcoding hard-bottom communities

Abstract

We developed a metabarcoding method for biodiversity characterization of structurally complex natural marine hard-bottom communities. Novel primer sets for two different molecular markers: the "Leray fragment" of mitochondrial cytochrome *c* oxidase, COI, and the V7 region of ribosomal RNA 18S were used to analyse eight different marine shallow benthic communities from two National Parks in Spain (one in the Atlantic Ocean and another in the Mediterranean Sea). Samples were sieved into three size fractions from where DNA was extracted separately. Bayesian clustering was used for delimiting molecular operational taxonomic units (MOTUs) and custom reference databases were constructed for taxonomic assignment. We found unexpectedly high values for MOTU richness, suggesting that these communities host a large amount of yet undescribed eukaryotic biodiversity. Significant gaps are still found in sequence reference databases, which currently prevent the complete taxonomic assignation of the detected sequences. Nevertheless, over 90% (in abundance) of the sequenced reads could be successfully assigned to phylum or lower taxonomical level. This identification rate might be significantly improved in the future, as reference databases are updated. Our results show that marine metabarcoding, currently applied mostly to plankton or sediments, can be adapted to structurally complex hard bottom samples, and emerges as a robust, fast, objective and affordable method for comprehensively characterizing the diversity of marine benthic communities dominated by macroscopic seaweeds and colonial or modular sessile metazoans, allowing for standardized biomonitoring of these ecologically important communities. The new universal primers for COI

48 can potentially be used for biodiversity assessment with high taxonomic resolution in a wide
49 array of marine, terrestrial or freshwater eukaryotic communities.

Introduction

Reliable methods for accurately and objectively assessing the biodiversity of marine environments are needed for a good understanding of these key ecosystems (Costello et al. 2010) in order to establish biodiversity baselines and monitor long-term biodiversity changes (Knowlton & Jackson 2008). Among marine ecosystems, shallow benthic hard-bottom communities are frequently considered to support the highest values of diversity, being arguably the most diverse ecosystems in the biosphere (Reaka-Kudla 1997; Agardy et al. 2005). Their proximity to humans place them among the best studied and most heavily impacted of all marine biomes. They are also the most influential for human ecology and economy. However, marine ecologists still lack robust, standardized tools for comprehensively surveying these communities.

An exhaustive analysis of these biomes by traditional morphological methods is impracticable due to their high complexity, the colonial or modular morphology of many groups and the abundance of tiny epibiotic forms (Mikkelsen & Cracraft 2001; Wangensteen & Turon 2017). In most instances, morphological surveys are limited to macro-organisms, and are often focused on a few taxonomic groups, strongly conditioned by the availability of taxonomic expertise. The taxonomic impediment (Wheeler et al. 2004) and the occurrence of cryptic species complexes (Knowlton 1993) further hinder the practicability of morphology-based methods.

In the last few years, the development of metabarcoding techniques, whereby thousands of species present in a given environmental sample can be detected by high-throughput sequencing and identified using molecular databases (Hajibabaei et al. 2011; Taberlet et al. 2012), has revolutionized biodiversity assessment. Metabarcoding approaches have been successfully used to characterize marine communities in relatively homogeneous substrates such as seawater (e.g. de Vargas et al. 2015; Chain et al. 2016) or marine sediments (e.g. Chariton et al. 2010; Fonseca et al. 2014; Pawlowski et al. 2014; Guardiola et al. 2015; Lejzerowicz et al. 2015) containing mostly small-sized organisms. Leray & Knowlton (2015) introduced methods for analysing the community DNA extracted from organisms collected in autonomous reef monitoring structures (ARMS) using COI metabarcoding. These artificial collectors have been used to analyse other genetic markers such as 18S (Pearman et al. 2016). However, metabarcoding methods have been scarcely used to characterize complex communities dwelling on marine natural hard-bottom substrates. These environments pose new challenges related to sample treatment (given the

orders-of-magnitude variation in organisms' sizes) and to the lack of reliable nearly-universal primer sets, capable of amplifying the wide array of taxonomic groups inhabiting these communities.

In the present work, we introduce a metabarcoding protocol for characterizing complex communities inhabiting natural marine hard substrates. The suitability and robustness of our methods are assessed by comparing the results from two independent universal eukaryotic molecular markers: a fragment of the multiple-copy nuclear gene for the small subunit of the ribosomal RNA (18S) and a fragment of the cytochrome *c* oxidase subunit I mitochondrial gene (COI). A multigene metabarcoding approach has been advocated to overcome limitations inherent to single marker studies (Drummond et al. 2015, Coward et al. 2015). The enhanced taxonomic resolution of COI (Tang et al. 2012) is partly counteracted by the lack of universality of most primer sets used for COI amplification, which may fail to amplify some eukaryotic taxa (Deagle et al. 2014). To overcome this problem, we introduce a new primer set, featuring a high ratio of degenerate positions, designed for enhancing universality in the amplification of the "Leray fragment" (Leray et al. 2013) of COI in most eukaryotic groups.

We tested this methodology by studying eukaryotic biodiversity patterns on eight different shallow benthic communities (fig. 1) sampled from two marine national parks in Spain (one in Western Mediterranean, and another in Northeastern Atlantic). Size fractionation has been proposed as a necessary step in metabarcoding when organisms on a sample have unequal biomass (Elbrecht et al. 2017). This procedure allowed recovery of 30% more taxa of freshwater invertebrates than unsorted samples (Elbrecht et al. 2017). Size fractionation has been used in the marine environment for metabarcoding macrobenthos in sediment samples (e.g., Aylagas et al. 2016), mobile organisms in settlement plates (e.g., Leray & Knowlton 2015, Ransome et al. 2017), or zooplankton (e.g., Liu et al. 2017), but it has never been applied to samples with organisms spanning several orders of magnitude in size such as hard-bottom communities (from macrophytes to microbes). We sieved each sample into three size fractions, which corresponded to the distinction between mega-, macro- and meiobenthos (Rex & Etter 2010), a separation with important correlates in terms of structure and function of benthic communities (e.g., Warwick & Joint 1987, Galerón et al. 2000, Rex et al. 2006).

Total community DNA was extracted separately for each fraction, and each extract was then metabarcoded in parallel runs using 18S and COI markers. We also analyzed unsieved sediment

samples from a tidal lagoon for comparative purposes. Our main objective was to develop and apply a method for characterizing complex marine hard-substrate communities using DNA metabarcoding. To this end, we (1) tested the effects of size-fractionation in the detection of marine taxa spanning widely different sizes, (2) assayed a modified primer set for COI, (3) generated new reference databases, (4) compared the relative performance of 18S and COI in terms of taxonomic accuracy and biodiversity patterns obtained, and (5) generated baseline information for biodiversity assessment and biomonitoring of benthic communities in Marine Protected Areas.

Materials and Methods

Sampling

Samples were taken by scuba diving from different shallow hard-bottom communities inside two national parks in Spain: Cíes Islands (Atlantic Islands National Park, Galicia, Northeastern Atlantic, 42.22ºN, 8.90ºW) and Cabrera Archipelago National Park (Balearic Islands, Western Mediterranean, 39.13ºN, 2.96ºE). A map of sampling locations is shown in fig. S1. The rationale for the choice of the communities was to have the most representative habitats along a depth gradient of the rocky littoral of these national parks for the purpose of obtaining baseline inventories for future monitoring and management efforts. Atlantic communities were sampled in May 2014, from two different communities of photophilous algae (3-5 m deep), one dominated by *Cystoseira nodicaulis* (a) and another dominated by *Cystoseira tamariscifolia* (b), a sciaphilous community dominated by *Saccorhiza polyschides* (16 m deep) (c) and detritic rhodolith beds (maërl bottoms) (ca. 20 m deep) (d). Mediterranean communities were sampled in September 2014 from a photophilous algal assemblage dominated by *Lophocladia lallemandii* (e), a photophilous community with an heterogeneous algal composition (f) (both from 5-10 m deep), a sciaphilous precoralligenous community (30 m deep) (g), and detritic rhodolith beds (ca. 50 m deep) (h). Although rhodolith beds are not strictly rocky communities, they are included in this work because they share with them the three-dimensional complexity and much of the biodiversity present, as we sampled communities just adjacent to rocky slopes. These detritic communities were sampled (three replicates each) by using a cylindrical PVC corer with a diameter of 30 cm and a height of 5 cm. All other hard-bottom communities (three replicates each) were sampled by carefully scraping a 25x25 cm quadrat with chisel and hammer. All

144 samples were placed underwater inside polyethylene bags. Water was eliminated through a 63
145 µm mesh sieve shortly after sampling, being then replaced by 96% ethanol. The material retained
146 in the filter was washed back to the sample bag with ethanol. Samples were stored at -20 ºC upon
147 arrival to the laboratory, until further processing. For comparative analyses, three additional
148 samples from soft-bottom sandy sediments were taken using a corer (3.5 cm in diameter) during
149 the low tide from a tidal lagoon in Cíes Islands (Lago dos Nenos, 42.223ºN, 8.905ºW) and
150 preserved whole in 96% ethanol.

151

152 Sample pre-treatment and DNA extraction

153 The samples were separated into three size fractions (A: > 10 mm; B: 1 – 10 mm; C: 63 µm – 1
154 mm) using a column of stainless steel sieves (www.cisa.net), washing thoroughly under high-
155 pressure freshwater. All separated fractions were then recovered in 96% ethanol, homogenized
156 using a 600 W hand blender and stored at -20 ºC until DNA extraction. All equipment was
157 thoroughly washed and cleaned with diluted sodium hypochlorite between successive samples.
158 The three sediment samples from the lagoon were processed directly without any sieving, and
159 manually homogenized. For total DNA extraction, 10 g of each homogenized sample were
160 purified using PowerMax Soil DNA Isolation Kit (www.mobio.com). DNA concentration of
161 purified extracts was assessed in a Qubit fluorometer (www.lifetechnologies.com) and, if
162 needed, concentrated in a Speedvac system (www.thermoscientific.com) until DNA
163 concentration of > 5 ng/µl was achieved.

164

165 Reproducibility and negative controls

166 One of the homogenized samples (from the Atlantic community dominated by *Cystoseira*
167 *tamariscifolia*) was extracted in triplicate and amplified independently, in order to check the
168 reproducibility of the DNA extraction procedure. One of these extractions was then amplified
169 using three PCR reactions with different sample tags, in order to check the reproducibility of the
170 PCR amplification and the possible bias introduced by mismatches due to sample tags
171 (O'Donnell et al. 2016). The variability of these samples (due to random errors during the PCR
172 in the latter case, and due to the addition of random errors during the PCR plus the variability in
173 the DNA extraction procedure in the former case) was also compared with the natural ecological

variability assessed by the three different replicates obtained from the same community.

Two different kinds of negative controls were used during the process. A standard PCR-blank was amplified using the elution buffer of the DNA isolation kit as a sample. A negative control for the pre-treatment separation protocol was done by using a sand sample charred in a muffle furnace at 400 ºC for 24 h to remove all traces of DNA. This muffled sand was sieved and extracted using the same procedure used for the samples. 2 PCR-blanks and 2 negative controls were run alongside the samples in the same sequencing plates.

DNA amplification and library preparation

Two different metabarcoding markers were amplified: 18S and COI. For the V7 region of 18S rRNA, the recently developed 18S_allshorts primers were used: forward: 5'-TTTGTCTGSTTAATTSCG-3' and reverse: 5'-TCACAGACCTGTTATTGC-3 (Guardiola et al. 2015). These primers show a marked universality across eukaryotic groups (see *in silico* analysis and primer logos in Guardiola et al. 2015). To these primers, 8-base sample-specific tags were attached (the same tag at both ends in order to detect intersample chimeric sequences). The PCR conditions followed Guardiola et al. (2015), using a standardized amount of sample (10 ng of purified DNA per sample).

For the 5' region of COI, we used a new highly degenerated primer set (henceforth Leray-XT), which includes the reverse primer jgHCO2198 5'-TAIACYTCIGGRTGICCRAARAAYCA-3' (Geller et al. 2013) and introduces a novel forward primer mlCOIintF-XT 5'-GGWACWRGWTGRACWITITAYCCYCC-3', modified from the mlCOIintF primer (Leray et al. 2013) by incorporating two more wobble bases and two inosine nucleotides in the most degenerate positions, for increased universality across eukaryotic groups. This was done after manually checking the original primer against representative sequences of the main eukaryotic groups obtained from the Genbank database. Sample tags were attached to both ends of the primers as before. Amplification of COI used AmpliTaq Gold DNA polymerase, with 1 µl of each 5 µM forward and reverse 8-base tagged primers, 3 µg of bovine serum albumin and 10 ng of purified DNA in a total volume of 20 µl per sample. The PCR profile included a denaturing step of 10 min at 95 ºC, 35 cycles of 94 ºC 1 min, 45 ºC 1 min and 72 ºC 1 min and a final extension of 5 min at 72 ºC.

After PCR, quality of amplifications was assessed by electrophoresis in agarose gel. All PCR products were purified using Minelute PCR purification columns (www.qiagen.com) and pooled by marker (83 samples per marker, including blanks and replicates used for the reproducibility study). Two Illumina libraries were built from the DNA pools using the Metafast protocol at Fasteris SA (Plan-les-Ouates, Switzerland, www.fasteris.com). This protocol incorporates Illumina adapters using a ligation procedure without any further PCR step, thus minimising biases. Each library was sequenced independently in an Illumina MiSeq platform using v3 chemistry (2x150 bp paired-end run for 18S and 2x300 bp paired-end run for COI).

In silico evaluation of the new COI primer set

We tested *in silico* the coverage of the new primer set for COI and compared it with the original Leray set (Leray et al. 2013). This comparison was done for all metazoan phyla and for the rest of eukaryotic groups for which enough sequence information was available. A set of full sequences for COI was downloaded from Genbank and the ability of each primer set to amplify the different species was assessed using ecopcr and the ecotaxstat function (Ficetola et al. 2010), allowing 3 mismatches per primer. This approach cannot be used directly with partial COI sequences (corresponding to the standard barcoding region), since these barcodes lack the reverse primer binding sequence. Thus, for those eukaryotic groups where not enough complete COI sequences were available (Dinoflagellata, Rhodophyta and Stramenopiles), we ran ecopcr and ecotaxstat against sets of COI barcode sequences with an artificial jgHCO2198-matching sequence attached to the 3' end. This allowed us to test the coverages of the internal forward primers and to compare the performance of the new Leray-XT primer set to that of the original Leray primers, since both sets are sharing the reverse primer. Primer logos (Crooks et al. 2004) were also obtained to summarize conservation of primer sequences across all eukaryotic groups.

Metabarcoding pipeline

We based our metabarcoding pipeline on the OBITools software suite (Boyer et al. 2016). The length of the raw reads was trimmed to a median Phred quality score higher than 30, after which paired-reads were assembled using illuminapairedend. The reads with alignment quality scores higher than 40 were demultiplexed using ngsfilter. A length filter (obigrep) was applied to the

assigned reads (75 – 180 bp for 18S and 300 – 320 bp for COI). The reads were then dereplicated (using obiuniq) and chimeric sequences were detected and removed using the uchime_denovo algorithm implemented in vsearch (http://github.com/torognes/vsearch). The MOTUs were then delimited using the Bayesian clustering algorithm implemented in CROP (Hao et al. 2011). This algorithm results in variable thresholds for delimiting MOTUs across different branches of the taxonomic tree, following the natural organization of the clusters in multidimensional sequence space. The following parameter sets were used: l=0.3, u=0.5 for 18S (Guardiola et al. 2016) and l=1.5, u=2.5 for COI. These values were chosen to avoid overclustering of several species into single MOTUs (Wangensteen & Turon 2017).

The taxonomic assignment of the representative sequences for each MOTU was performed using ecotag (Boyer et al. 2016), which uses a local reference database and a phylogenetic tree-based approach (using the NCBI taxonomy) for assigning sequences without a perfect match. Ecotag searches the best hit in the reference database and builds a set of sequences in the database which are at least as similar to the best hit as the query sequence is. Then, the MOTU is assigned to the most recent common ancestor to all these sequences in the NCBI taxonomy tree. With this procedure, the assigned taxonomic rank varies depending on the similarity of the query sequences and the density of the reference database. For 18S, we used the db_18S_r117 reference database (Guardiola et al. 2015), obtained by *in silico* ecoPCR (Ficetola et al. 2010) with the 18S_allshorts primer set against the release 117 of the EMBL nucleotide database. This database includes 26,125 reference sequences from all major eukaryotic groups. For COI, we developed a mixed reference database by joining sequences obtained from two sources: *in silico* ecoPCR against the release 117 of the EMBL nucleotide database and a second set of sequences obtained from the Barcode of Life Datasystems (Ratnasingham & Hebert 2007) using a custom R script to select the Leray fragment. This newly generated database (db_COI_MBPK) included 188,929 reference sequences (March 2016) from a wide taxonomic range. Both reference databases are publicly available from http://github.com/metabarpark/Reference-databases.

After taxonomic assignment, the final refining of the datasets included taxonomic clustering of MOTUs assigned to the same species, minimal abundance filtering (unassigned MOTUs with less than 10 reads and assigned MOTUs with less than 5 reads were deleted), blank correction and abundance renormalization to remove spurious false positive results due to random tag switching (Wangensteen & Turon 2017). Since we were interested only in eukaryotic diversity,

all MOTUs assigned to prokaryotes or to the root of the Tree of Life were removed from the analyses. Samples having less than 10,000 reads in the final datasets, after all filtering procedures, were considered as failed and deleted from the analyses. The pipelines used for both metabarcoding markers are summarized in supplementary material, table S1.

Statistical analyses

All analyses were performed in R v 3.3.0 (https://www.R-project.org/). Package vegan (Oksanen et al. 2016) was used for rarefaction analyses (function rrarefy), calculations of Bray-Curtis dissimilarity matrices (function vegdist), comparison of these matrices (function mantel), and group representation in nMDS diagrams (functions isoMDS, ordiellipse and ordispider). Rarefaction analyses for α-diversity were carried out using a rarefaction size of 10,000 reads and 500 bootstrap replicates per sample. Package vioplot (Adler 2005) was used for plotting the results as violin plots and package dunn.test (Dinno 2017) was used to test differences in α-diversity between fractions. All calculations of Bray-Curtis dissimilarities were performed using fourth root-transformed abundance values of relative frequencies (normalized by dividing read abundances of each MOTU by the total reads for each sample). For assessing the reproducibility of the extraction and PCR procedures, weighted UniFrac distances between technical replicates were calculated using package phyloseq (McMurdie & Holmes 2013) and compared with the same distances obtained from ecological replicates. This method takes into account not only differences in abundances of the occurring MOTUs for calculating dissimilarities between samples, but also the phylogenetic distance between these MOTUs. For assessing the effect of size fractionation on the detectability of organisms with different sizes, we classified the MOTUs into the following ecological size-categories: (1) macroscopic seaweeds, (2) modular metazoa, (3) macrofaunal unitary metazoa, (4) meiofaunal metazoa, (5) microorganisms and (6) unassigned. The percentages of reads obtained for each category in the three size fractions were compared for each hard-bottom community and to the sediment samples from the tidal lagoon.

Results

In silico evaluation of the new COI primer set

The taxonomic coverage at the species level of the new Leray-XT primers and the original Leray

296 set is shown in fig. 2 for different groups of eukaryotic and metazoan phyla. The coverage of the
297 new Leray-XT primers was enhanced for most eukaryotic groups compared to the original Leray
298 set, and reached values higher than 96% for all major groups analyzed, except for Fungi,
299 Viridiplantae and Ciliophora. Values of 96% or higher were also reached for all major metazoan
300 phyla, except for Cnidaria, Platyhelminthes and Porifera. Even for these three phyla, there was a
301 substantial coverage enhancement respect to the original Leray set (59% to 89% for Cnidaria,
302 35% to 82% for Platyhelminthes and 67% to 89% for Porifera). The combined coverage for all
303 Metazoa with the Leray-XT primers was 96.1%, compared with 86.0% for the original Leray set.
304 The primer logos for a combined set of 149,908 COI sequences from 38 different phyla of all
305 eukaryotic lineages are shown in fig. S2.

306

307 *Taxonomic summary of the new reference databases*

308 The number of sequences from different taxonomic groups included in the reference databases
309 used for our analyses is summarized in table S2. Although the total number of different reference
310 sequences for COI is one order of magnitude higher than for 18S, some important taxonomic
311 groups are remarkably absent from the COI reference database, such as Choanozoa, Foraminifera
312 or several fungal phyla, while others are scarcely represented, such as Cercozoa, Excavata or
313 Cryptophyta. Among groups with macro-organisms, the low representation of Viridiplantae in
314 the COI database is noteworthy, while Chordata are poorly represented in the 18S database
315 (1.3% of the total sequences vs 21.2% in COI).

316

317 *Sequencing depth and α-diversity patterns*

318 We metabarcoded a total of 83 samples (36 subsamples from 4 benthic communities in Cíes, 36
319 from 4 communites in Cabrera, 3 samples from a tidal lagoon, 4 additional samples used for
320 studying reproducibility, 2 blanks and 2 negative controls). After the refining procedures, our
321 final dataset for 18S comprised a total of 8,266,952 reads, with an average of 105,987 reads per
322 sample (range: 61,828 – 190,046). For COI, our final dataset included 10,093,453 reads, with an
323 average of 134,580 reads per sample (range: 10,154 – 423,822). One sample from the 18S
324 dataset and four samples from the COI dataset were removed from the analyses due to low
325 number of reads (< 10,000). Controls had a negligible number of reads.

The number of total MOTUs detected from all samples by Bayesian clustering was 5,067 for 18S, from which 4,130 (81.5%) could be assigned to the level of phylum or lower. These assigned MOTUs accounted for 98.1% of the total 18S reads. As expected, the number of MOTUs yielded by COI from the same samples was higher: 21,452, from which 12,369 (57.6%) could be taxonomically assigned to the level of phylum or lower. The assigned MOTUs accounted for 91.3% of total COI reads. Our final datasets, including the sequences of all MOTUs, their taxonomic assignment and their abundances in each sample, are presented as supplementary material (tables S3 and S4) and are available from the Mendeley data repository (https://data.mendeley.com/datasets/nm2c97fjng/1).

The different fractions of the sampled communities showed similar patterns of α-diversity for 18S and COI after rarefaction (fig. 3). Using either marker, a trend can be observed whereby larger fractions (A and B) had similar values for MOTU richness, whereas the smallest fraction (C) was significantly more diverse than the other two (all Kruskal-Wallis followed by Dunn's tests $p<0.01$) with the exception of the Mediterranean detritic rhodolith community, where no differences in α-diversity between fractions were detected using COI (Kruskal-Wallis $p=0.14$). The α-diversity values for the lagoon sediment samples were in the same range than those of samples from fractions C of hard substrate communities. Values of α-diversity measured from COI were roughly three times higher than values obtained from 18S.

Taxonomic assignment and database gaps

The number of taxa identified at phylum or lower categories for both markers is shown in fig. 4. A clear trend emerges: the lower the category, the less coincidence between the taxa found with both markers. Thus, at the phylum level, 87.5% of the phyla detected with COI were also recovered with 18S. The corresponding figures were 81.4% for Class, 66.2% for Order, 42.6% for Family, 24.5% for Genus, and 6.5% for species-level taxa. Moreover, 18S detected a higher number of taxa in the different categories, except for species: 803 species were detected with COI vs 615 with 18S.

The numbers of MOTUs detected by phylum (fig. 5) showed that both markers, COI and 18S, were able to detect those groups composed of medium- or big-sized organisms, such as major metazoan phyla or macroscopic seaweeds. The detection of groups comprising microscopic

organisms was usually more reliable using 18S than COI. For example, 19 metazoan phyla could be detected in our samples using COI, while the 18S assignment detected these same 19 phyla plus the microscopic Kinorhyncha, Loricifera and Gnathostomulida. Due to remarkable gaps in the reference database (as seen in table S2), our assignment procedure for COI was unable to identify any sequence from microscopic groups such as Apusozoa, Choanozoa, Heliozoa, Protalveolata or Rhizaria (including Foraminifera, Cercozoa and Radiozoa), which could be detected by 18S. However, COI was able to detect and distinguish a higher number of MOTUs than 18S for most macroscopic phyla. Moreover, the assignment at the species level was more reliable using COI than 18S. An assignment with an identity percent higher than 97% using COI leads in general to correct species identification; whereas, in many cases, the assignment of 18S by the ecotag algorithm (even at 100% identity) yielded taxa not present in the studied areas. This happens because related species included in the reference database share exactly the same sequence for the 18S fragment used, whereas cases of synonymous sequences for different species are extremely rare using COI. Although errors in taxonomic annotation in the databases can also affect species identification, such errors would be present for both markers. MOTUs with high abundance of reads could be in general identified to the species level using COI, whereas they were often identified to higher taxonomic ranks using 18S (fig. S3). Unassigned MOTUs are those with least abundances, using either marker (fig. S3).

Patterns of MOTU abundances

The abundances of reads assigned to major eukaryotic groups at a level of phylum or lower are presented (in percentages) in fig. 6 for COI and 18S (the same information split by replicate samples is presented in fig. S4). The rates of unassigned sequences were, in all cases, higher for COI than for 18S. The unassigned reads were always most abundant in the smallest fraction (fraction C) of each community and were particularly abundant in the lagoon sediment samples (average: 35.7% of COI reads vs 19.7% of 18S reads). Sequences identified as small Metazoa such as Annelida or Arthropoda were also clearly more abundant in the smallest fractions, whereas big macroscopic seaweeds such as Rhodophyceae or Phaeophyceae tended to be dominant in the biggest fractions (A and B). Colonial and modular Metazoa such as Porifera, Cnidaria or Bryozoa were distributed across all fraction sizes. The lagoon sediment samples (unsieved) were enriched in microscopic organisms such as Bacillariophyta (29.4% of COI

reads), Oomycetes (2.9% of COI reads) or Ciliophora (6.0% of 18S reads), which were scarce in the sieved samples (averages of 0.75%, 0.37% and 0.20% respectively). Although some differences may be observed between both markers (e.g.: higher abundance of reads of Mollusca and Porifera from 18S and more reads of Arthropoda and Rhodophyta from COI), the overall patterns of read abundances were similar for 18S and COI. The three ecological replicates per community were also similar in composition (fig. S4).

The number of MOTUs assigned to the different phyla in each sample are shown (in percentages) in fig. 7 for COI and 18S (the same information split by replicate samples is presented in fig. S5). Compared to the read abundances of fig. 6, a higher dominance of small-sized MOTUs is apparent. The percentages of MOTUs assigned to microeukaryotes were notably higher for 18S than for COI. So, the sum of MOTUs assigned to ciliates, dinoflagellates, Bigyra and other protists accounted for 15.11% of assigned 18S MOTUs, while this sum was just 2.27% of the assigned COI MOTUs. Other groups with higher relative richness measured by 18S were annelids (9.56% of 18S MOTUs vs 5.70% of COI MOTUs), nematodes (3.38 vs 0.48%) and flatworms (2.98 vs 0.21%). In contrast, COI detected relatively more MOTUs than 18S for rhodophytes (19.11% COI vs 12.41% 18S), cnidarians (14.9% vs 9.13%), arthropods (15.07% vs 11.16%), mollusks (8.93% vs 5.25%), oomycetes (4.11% vs 0.59%) and diatoms (7.13% vs 4.38%). All other groups differed in less than 2% of total assigned MOTUs among markers. Again, the three replicates per community showed a similar composition in terms of MOTU richness per phylum (fig. S5).

Reproducibility

Weighted UniFrac dissimilarity indices calculated among PCR replicates and among extraction replicates were compared with dissimilarities among ecological replicates. For 18S, the average dissimilarity between PCR-replicates was 0.0003 ± 0.0001, smaller than between DNA extractions from the same sample (0.0016 ± 0.0016) and two orders of magnitude smaller than between ecological replicates (0.0390 ± 0.0120). For COI, the equivalent values were 0.102 ± 0.017, 0.117 ± 0.017 and 0.283 ± 0.016 for PCR-replicates, extraction-replicates and ecological replicates, respectively. Thus, technical replicates yielded significantly more similar results than ecological replicates, indicating the robustness of the protocols. Pie charts representing the read

abundances of major groups detected at the different levels of replication are shown in fig. S6, which highlight the differences in the relative abundances of MOTUs among ecological replicates compared to extraction replicates and PCR replicates, with both markers.

Ordination patterns of community structure

Non-metric multidimensional scaling (nMDS) plots showing the ordination of the studied communities are shown in fig. S7 for COI and 18S. Similar ordination patterns were recovered from both markers, and the two Bray-Curtis matrices (18S and COI) were highly correlated (mantel test, r=0.897, p<0.001). Samples from the three fractions of each community grouped together, with overlap of the inertia ellipses in many cases. Samples from both photophilous atlantic communities clustered together, and the same applies to both photophilous mediterranean communities, suggesting the presence of a high proportion of shared MOTUs between these shallow communities. On the other hand, mediterranean and atlantic samples appeared separated, and a gradient from shallower (well-lit photophilic communities) to deeper, sciaphilous communities was apparent. Samples from the shallow lagoon appeared as a tight cluster, well-separated from other benthic communities.

Effect of size fractionation in the detectability of MOTUs

Venn diagrams representing the MOTUs detected in the three fractions are presented in fig. 8. There is an important overlap with 18S (73% of MOTUs were detected in the three fractions), while this overlap is substantially reduced with COI (56% of MOTUs). In addition, fraction C of COI has twice as many unique MOTUs as fraction C of 18S, and more than one quarter of COI MOTUs (27%) are found exclusively in the two smaller fractions (B and C).

Differences between fractions were more evident in terms of read abundance than presence/absence of MOTUs. The percentages of read abundances belonging to different ecological size-categories recovered from every fraction of the analyzed communities are presented in fig. 9 for COI and 18S. The microorganismal category was recovered mainly in the unsieved samples from the lagoon, which were mostly composed of meiofauna, microorganisms and unassigned reads. Interestingly, most reads of microorganisms detected in fractions A, B and C of hard-bottom communities, belonged to *Symbiodinium sp.* or *Amphidinium sp.*,

dinoflagellates which are symbionts of macrofaunal anthozoans. As expected, macroalgae were more abundant in fractions A and B, whereas meiofaunal reads were more abundant in fractions C. Reads of modular metazoans were more evenly distributed among the three fractions.

Discussion

The application of metabarcoding techniques to characterize marine hard bottom communities has been hindered by a lack of standardized methods for sample collection and treatment, the scarcity of universal primers capable of amplifying the wide array of taxonomic groups present in these communities and the need of bioinformatic procedures able to cope with the high degree of genetic diversity obtained. We think that the procedures presented here, which include extraction of DNA from separate size fractions, a novel set of highly degenerate primers for COI, capable of amplifying most eukaryotic groups, and improved bioinformatic pipelines for data treatment including new reference databases for Eukarya, allow to overcome many of the challenges related to metabarcoding of structurally complex macroscopic benthic communities. In this work, we tested this approach on the eukaryotic diversity present in eight ecologically diverse littoral benthic communities. These procedures have already proven useful to detect effects of three invasive algae on the small-sized organisms of littoral communities in a different set of samples (Wangensteen et al. in press) and can be applied, with the necessary adjustments, for biodiversity assessment in a wide array of marine, terrestrial or freshwater eukaryotic communities.

Sample pre-treatment, the benefits of size fractionation

The partitioned metabarcoding of size fractions filtered through a column of sieves allows characterization of structurally complex communities at different levels, which would be impossible using whole samples, due to the high number of DNA copies from organisms of bigger biomass outnumbering the smaller ones and hampering their detection (Coward et al. 2015, Elbrecht et al. 2017). We have shown that the smallest fractions are the most diverse (fig. 3) and are enriched in meiofaunal elements, which can be detected because most of the biomass

from big-sized organisms is retained within fractions A and B (fig. 9). Even if there is an important qualitative overlap (fig. 8), many MOTUs appear in some abundance only in fraction C. Thus, without fraction C, not only the exclusive MOTUs disappear, but 280 MOTUs of 18S (5.5% of the total) and 3,317 MOTUs of COI (15.4%) were left with less than 5 reads. These MOTUs would have been removed during the minimal filtering clean-up without the contribution of reads from the smallest fraction.

An additional advantage of this procedure is the removal of a significant fraction of microorganisms (prokaryotes and the smallest microeukaryotes), together with most of the extra-organismal DNA in the form of small remains, cell debris, or extracellular DNA (Creer et al. 2016), which are not retained in the last sieve (63 μm). Microeukaryotes are known to be genetically diverse and under-represented in genetic databases, which introduces problems during bioinformatic analyses, particularly for clustering and taxonomic assignment algorithms. They are better removed from the samples by sieving whenever they are not the main study target. The improved results in the assignment can be appreciated by the higher rate of unassigned reads from the tidal lagoon (unsieved) compared to sieved samples from littoral communities (figs. 6 and 9). Moreover, most MOTUs with high read abundances could be assigned to the species level using COI (fig. S3), while unassigned MOTUs were typically the least abundant, suggesting again the reference database bias towards big and abundant species. Therefore, in studies mainly aimed at characterizing macro- and meio-benthic components, some physical filtering step is advisable during sample pre-treatment. Size-fractionation has been used to separate relevant compartments in metabarcoding studies of planktonic organisms (e.g., Logares et al. 2014, Massana et al. 2015, Liu et al. 2017) and in some studies of sedimentary bottoms (e.g. Chariton et al. 2010, Coward et al. 2015, Aylagas et al. 2016). However, this point had not been addressed for hard-bottom benthic communities, where size-differences encompass many orders of magnitude. The closest reference is the study of artificial settlement surfaces (ARMS, Leray & Knowlton 2015, Ransome et al. 2017) where organisms were separated into sessile biota (processed in bulk) and three size-classes of motile organisms, being the smaller fractions the most diverse. Size-fractionation should be mandatory for adequate recovery of biodiversity in hard substrate benthic communities.

Choice of a proper metabarcoding marker, COI vs 18S

The amplification of COI resulted in more MOTUs than 18S (by a factor of 4) and more resolving power at the species level (803 vs 615 species-level assignments), at the cost of a higher proportion of unassigned MOTUs (overall 42.4% and 18.5%, respectively), a result consistent with previous findings (e.g., zooplankton Clarke et al. 2017). The use of COI as a metabarcoding marker has been criticized in the past, arguing that high rates of sequence variability impair the design of truly universal primers and hamper the bioinformatic analyses (Deagle et al. 2014), but attempts have been made recently to incorporate COI data in metabarcoding studies (e.g., Leray and Knowlton 2015, Berry et al 2015, Aylagas et al. 2016, Elbrecht & Leese, 2017). . However, COI presents two major advantages compared to other possible markers. First, the steadily growing international effort to develop a public DNA barcoding database with curated taxonomy, which vastly facilitates taxonomic assignment. The BOLD database (Hebert et al. 2003; Ratnasingham & Hebert 2007), based mainly in COI barcoding, currently includes over 4 million sequences belonging to more than 500,000 different species, curated and identified by expert taxonomists. It is highly unlikely that any comparable effort might be undertaken for any other marker in the next future. Second, the high mutation rate of COI practically ensures unequivocal identification at the species level, whereas the highly conserved sequence of 18S makes it usually impossible to distinguish at the genus or family levels, or even at higher ranks (Tang et al. 2012). Species-level resolution is crucial for calculating ecological indices or detecting non-native species (Comtet et al. 2015, Aylagas et al. 2016).

Overall, then, we favour the use of COI amplicons for characterizing complex marine communities. The use of 18S can be recommended only when the information at the species level is not crucial. For example, to assess overall impacts related to human activities such as fisheries, aquaculture or mining facilities. In this case, the impact may be expected to affect abundances and composition at higher taxonomical levels. Studying these impacts using 18S may benefit from the less computationally demanding and faster bioinformatics processing of 18S data than that of COI data.

Our universal Leray-XT primer set for COI features high values of *in silico* coverage (fig. 2) and was able to successfully amplify a wide range of eukaryotic organisms belonging to 19 phyla of Metazoa and all major marine lineages of Eukaryota (fig. 5). The undetectability of some lesser groups in this study is more related to the incompleteness of reference databases (table S2),

rather than to PCR failure of our COI primer set. We recommend the use of this universal primer set (mlCOIintF-XT and jgHCO2198) for COI metabarcoding analyses of marine samples or other environmental or community DNA projects, especially when a wide taxonomic range of eukaryotes is expected and species-level resolution is necessary. We note, however, that these primers have limited ability for detecting some groups (e.g., Viridiplantae and Ciliophora, fig. 2, table S2). Thus, specific primers or a different marker should be used if these taxonomic groups constitute the main study target.

It is remarkable that ordination analyses of our data yielded robust and comparable results, disregarding the marker chosen (fig. S7). The two distance matrices were highly correlated, indicating that the same general ecological information is retrieved with both markers. This implies that robust and objective methods for impact studies or comparisons among communities may be designed and implemented with different markers.

Estimates of α-diversity: a comparison with morphological studies.

The ultimate aim of metabarcoding is to objectively determine which species are present in a given environmental sample. The values obtained for eukaryotic richness in this study are astoundingly large. Using 18S, 4,203 different MOTUs were detected in the four mediterranean communities combined, and 3,914 MOTUs in their atlantic counterparts. The respective values for COI MOTU richness are 14,092 for Cabrera and 13,708 for Cíes. These values are comparable to 17,000, which is the total number of morphological species described for the whole Mediterranean Sea (Coll et al. 2010). Rarefaction to just 10,000 reads per sample (fig. 3) yielded values for MOTU richness of roughly 200-600 MOTUs per replicate using 18S or 500-1500 MOTUs per replicate using COI. However, an adequate benchmarking of these values against richness detected with traditional (morphology-based) techniques for this kind of communities is still lacking. For sediment macroinvertebrates, Aylagas et al. (2016) showed that the Leray fragment generated over 50% of matches (depending on the protocol and lab conditions tested) between morphologically and molecularly inferred taxonomic composition. Biotic indices based on both sources were also well correlated. We cannot perform such a direct comparison with our samples as morphological information is not available. However, we can draw upon published studies of the same communities analyzed to gain an idea of the relative

performance of metabarcoding for characterizing biodiversity in hard substratum communities.

Traditional methods for community characterization in these communities rely on randomly allocating standardized sampling units (usually quadrats of 20x20 or 25x25 cm) and either collecting the biota through scraping, performing in situ visual censuses, or analyzing photographs. A comparison of the three methods was made precisely on the Cabrera Archipelago (Sant et al. 2017), highlighting relative differences in information and cost/benefits among methods. However, even the best performing method (scraping) identified a total of 262 species, an order of magnitude lower than we obtained in Cabrera from just the coarser fraction (A): 3,085 MOTUs with 18S and 9,333 with COI.

Other studies have analyzed species richness of the macrofauna and macroflora in both National Parks or geographically close areas in communities identical or similar to the ones studied here. In addition, a monograph is available on the taxonomy of benthic groups in Cabrera Archipelago (Alcover et al. 1993). In table 1 we have collated the information from these works and compared richness values with the ones obtained in our study of the corresponding communities. Metabarcoding largely outperforms morphological inventories, detecting on average 3.16 and 8.88 times more MOTUs (18S and COI, respectively) than reported in exhaustive morphological studies. Only in the case of Chlorophyta with both markers, Echinodermata of Cabrera with both markers, Mollusca with 18S and Phaeophyceae in the detritic of Cabrera with 18S did we detect a lower number of MOTUs than morphospecies reported. We must keep in mind that published results are often compilations of several works, while we have results for only a handful of samples taken at a single time point. The dominant genera and species mentioned in quantitative studies are in agreement with the results obtained with metabarcoding (tables S3 and S4).

Our results show that genetic estimates for diversity (especially those obtained from COI metabarcoding) largely exceed the results from morphological assessments, in agreement with other metabarcoding studies which reported unexpectedly higher genetic than morphological diversity estimates in comparable samples (Cowart et al. 2015), suggesting the existence of a large number of yet undescribed marine taxonomic lineages. Overall, then, metabarcoding seems well suited for biodiversity detection in hard bottoms, with the added advantage that it can target not just macro-organisms as most previous morphological studies did, but also meio- and micro-organisms. A more precise benchmarking, analyzing the same samples with both methods, remains to be performed in future studies.

601

602 *Taxonomic assignment. Current gaps in databases*

603 We did not add any new sequences to build our custom reference database. Instead, we
604 deliberately used only sequences already available from public repositories in order to assess the
605 completeness of current barcoding databases for marine taxa. Release r117 of the EMBL
606 repository was searched using *in silico* ecoPCR (Ficetola et al. 2010) with our metabarcoding
607 primer sets in order to obtain the reference sequences for our 18S and COI reference databases.
608 Our COI reference database was then enriched by adding sequences available from the BOLD
609 database (Ratnasingham & Hebert 2007). The rates of unassigned sequences in our results
610 suggest that important gaps still exist for both markers in the genetic repositories, which would
611 prevent the detailed identification of many marine organisms, in agreement with the concerns
612 expressed by other authors (Leray & Knowlton 2016). If a fine taxonomic identification of the
613 obtained sequences is desired for a given metabarcoding project, it is advisable not to rely
614 exclusively in the public repositories and obtain custom databases, including the generation of
615 sequences for known local species absent from the repositories. For many ecological
616 applications, however, it suffices that a particular MOTU is defined, its patterns of distribution
617 and abundance assessed, and changes over time monitored, even if a scientific name for that
618 MOTU is yet unavailable (Cordier et al. 2017). Moreover, the sequences of all MOTUs
619 (identified or not) detected by metabarcoding will remain in public repositories, so that
620 unidentified MOTUs might well be assigned a name in the future, as databases improve. The
621 same can hardly be said of morphological studies, where many taxa cannot be identified to
622 species level either, and are left inventoried under general names (e.g., “Nematoda spp or spX”)
623 without descriptions. Therefore, unlike metabarcoding datasets, currently used morphological
624 inventories contain a great deal of untraceable information that can never be used by other
625 researchers at any other place or time.

626 Database gaps affect the metazoan groups differentially. For example, despite being abundant
627 and diverse in benthic ecosystems, most bryozoans and cnidarians could be rarely assigned using
628 COI below the class or order level, whereas species of echinoderms, decapods or vertebrates
629 were usually successfully identified. A trend can also be seen of smaller sized groups, such as
630 Nematoda, Rotifera or benthic Copepoda, being left out of the databases, whereas bigger sized or
631 commercially important animals such as fish or decapods are well represented. There is no doubt

that taxonomic assignment of COI metabarcoding data will be more accurate and detailed in the future, as reference databases are populated by international barcoding initiatives, such as the Census of Marine Life (http://www.coml.org) or the Marine Barcode of Life (MarBOL, http://www.marinebarcoding.org). We strongly advocate for the continued public support and funding of such collaborative DNA-barcoding projects as a necessary tool towards the implementation of reliable and objective metabarcoding techniques for environmental assessment.

In this article, we showed how complex communities with organisms of a wide range of sizes can be tractable with an adapted community-DNA metabarcoding approach, even without the availability of taxonomic expertise, thus expanding the range of applications of this technique to ecosystems of enormous ecological and economic importance. At the same time, we have generated the first metabarcoding inventories for natural hard substrate communities, using Marine Protected Areas as our sampling settings, thus providing baseline information for future conservation-oriented research.

Conclusions

In this work we develop and apply a metabarcoding protocol for complex natural marine hard-bottom communities. Size fractionation is mandatory to adequately capture information for organisms of a range of sizes spanning several orders of magnitude. We assayed a novel primer set for the amplification of the "Leray fragment" of COI, introducing more degenerate positions for increased universality, as shown in *in silico* tests. Results show that COI recovers four times more diversity (in MOTU richness) than sequences of the ribosomal 18S molecule (v7 region). Reference databases are generated from publicly available sequences, showing that significant gaps still prevent complete taxonomic assignment of the sequences. Notwithstanding, assigned (at the phylum level) MOTUs represented >90% of reads for both markers. Our results show that marine metabarcoding, currently applied mostly to plankton or sediments, can be adapted to characterize the bewildering diversity of marine benthic communities dominated by macroscopic seaweeds and colonial or modular sessile metazoans. We used as a case study representative

sublittoral communities National Parks, thus generating baseline inventories for future biomonitoring of these communities with conservation and management implications.

Acknowledgements

We are indebted to the staff of the Atlantic Islands of Galicia and Cabrera Archipelago National Parks for sampling permits and invaluable logistic help. Thanks also to Xavier Roijals for his skilful bioinformatic assistance in the CEAB computing cluster facilities. We dedicate this work to Alex Macía, owner of the Cíes Diving club, who provided logistics in the Atlantic Islands, and who recently died while following his passion for diving.

References

Adler D (2005) vioplot: Violin plot. R package version 0.2. http://CRAN.R-project.org/package=vioplot.

Agardy T, Alder J, Dayton P *et al.* (2005) Coastal systems. In: Reid W, editor. Millennium ecosystem assessment: ecosystems and human well-being. Washington: Island Press. p. 513–549.

Alcover JA, Ballesteros E, Fornós JJ (1993) Història Natural de l'Archipèlag de Cabrera. Palma de Mallorca, Spain. Editorial Moll-CSIC.

Altaba CR (1993) Els mol.luscs marins: catàleg preliminar. In: Alcover JA, Ballesteros E, Fornós JJ, editors. *Història Natural de l'Archipèlag de Cabrera*. Palma de Mallorca, Spain. Editorial Moll-CSIC, 503-530.

Aylagas E, Borja A, Irigoien X, Rodríguez-Ezpeleta N (2016) Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. *Frontiers in Marine Science,* **3**, 96.

Ballesteros E (1993) Algues bentòniques i fanerògames marines. In: Alcover JA, Ballesteros E, Fornós JJ, editors. *Història Natural de l'Archipèlag de Cabrera*. Palma de Mallorca, Spain. Editorial Moll-CSIC, 503-530.

Ballesteros E (1994) The deep-water *Peyssonnelia* beds from the Balearic Islands (Western Mediterranean). *PSZNI Marine Ecology*, **15**, 233-253.

Berry O, Bulman C, Bunce M, Coghlan M, Murray DC, Ward RD (2015). Comparison of morphological and DNA metabarcoding analyses of diets in exploited marine fishes. *Marine Ecology Progress Series*, **540**, 167–181.

697 Boyer F, Mercier C, Bonin A *et al.* (2016) obitools: a unix-inspired software package for DNA
698 metabarcoding. *Molecular Ecology Resources*, **16**, 176–182.

699 Chain FJJ, Brown EA, MacIsaac HJ, Cristescu ME (2016) Metabarcoding reveals strong spatial
700 structure and temporal turnover of zooplankton communities among marine and freshwater ports.
701 *Diversity and Distributions,* **22**, 493–504.

702 Chariton AA, Court LN, Hartley DM *et al.* (2010) Ecological assessment of estuarine sediments
703 by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment*, **8**,
704 233–238.

705 Clarke LJ, Beard JM, Swadling KM, Deagle BE (2017) Effect of marker choice and thermal
706 cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution*, **7**, 873-
707 883.

708 Coll M, Piroddi C, Steenbeek J *et al.* (2010) The biodiversity of the Mediterranean Sea:
709 estimates, patterns, and threats. *PLoS One*, **5**, e11842.

710 Comtet T, Sandionigi A, Viard F, Casiraghi M (2015) DNA (meta)barcoding of biological
711 invasions: a powerful tool to elucidate invasion processes and help managing aliens. *Biological*
712 *Invasions,* **17**, 905–922.

713 Corbera J, Ballesteros E, Garcia Ll (1993) Els crustacis decàpodes. In: Alcover JA, Ballesteros
714 E, Fornós JJ, editors. *Història Natural de l'Archipèlag de Cabrera*. Palma de Mallorca, Spain.
715 Editorial Moll-CSIC, 579-587.

716 Cordier T, Esling P, Lejzerowicz F *et al.* (2017) Predicting the ecological quality status of
717 marine environments from eDNA metabarcoding data using supervised machine learning.
718 *Environmental Science & Technology*, **51**, 9118-9126.

719 Costello MJ, Coll M, Danovaro R *et al.* (2010) A census of marine biodiversity knowledge,
720 resources, and future challenges. *PLoS One*, **5**, e12110.

721 Cowart DA, Pinheiro M, Mouchel O *et al.* (2015) Metabarcoding is powerful yet still blind: a
722 comparative analysis of morphological and molecular surveys of seagrass communities. *PLoS*
723 *One*, **10**, e0117562.

724 Creer S, Deiner K, Frey S *et al.* (2016) The ecologist's field guide to sequence-based
725 identification of biodiversity. *Methods in Ecology and Evolution*, **7**, 1008–1018.

726 Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator.
727 *Genome Research,* **14**, 1188–1190.

728 Deagle BE, Jarman SN, Coissac E *et al.* (2014) DNA metabarcoding and the cytochrome c
729 oxidase subunit I marker: not a perfect match. *Biology Letters*, **10**, 20140562.

730 Dinno A (2017) dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums. R package
731 version 1.3.4. http://CRAN.R-project.org/package=dunn.test.

732 Drummond AJ, Newcomb RD, Buckley TR *et al.* (2015) Evaluating a multigene environmental
733 DNA approach for biodiversity assessment. *GigaScience*, **4**, 46.

734 Elbrecht V, Leese F (2017) Validation and development of COI metabarcoding primersfor
735 freshwater macroinvertebrate bioassessment. Frontiers in Environmental Science **5**, 11.

736 Elbrecht V, Peinert B, Leese F (2017) Sorting things out-assessing effects of unequal specimen
737 biomass on DNA metabarcoding. *Ecology and Evolution,* **7**, 6918-6926.

738 Ficetola GF, Coissac E, Zundel S *et al.* (2010) An in silico approach for the evaluation of DNA
739 barcodes. *BMC Genomics*, **11**, 434.

740 Fonseca VG, Carvalho GR, Nichols B *et al.* (2014) Metagenetic analysis of patterns of
741 distribution and diversity of marine meiobenthic eukaryotes. *Global Ecology and Biogeography*,
742 **23**, 1293–1302.

743 Galéron J, Sibuet M, Mahaut ML, Dinet A (2000) Variation in structure and biomass of the
744 benthic communities at three contrasting sites in the tropical Northeast Atlantic. *Marine Ecology*
745 *Progress Series*, **197**, 121-137.

746 Geller J, Meyer C, Parker M, Hawk H (2013) Redesign of PCR primers for mitochondrial
747 cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic
748 surveys. *Molecular Ecology Resources* **13**, 851–61.

749 Gili JM, Garcia-Rubies A, Tur JM (1993) Els cnidaris bentònics. In: Alcover JA, Ballesteros E,
750 Fornós JJ, editors. *Història Natural de l'Archipèlag de Cabrera*. Palma de Mallorca, Spain.
751 Editorial Moll-CSIC, 549-559.

752 Guardiola M, Uriz MJ, Taberlet P *et al.* (2015) Deep-sea, deep-sequencing: Metabarcoding
753 extracellular DNA from sediments of marine canyons. *PLoS One*, **10**, e0139633.

754 Guardiola M, Wangensteen OS, Taberlet P *et al.* (2016) Spatio-temporal monitoring of deep-sea
755 communities using metabarcoding of sediment DNA and RNA. *PeerJ*, **4**, e2807.

756 Hajibabaei M, Shokralla S, Zhou X et al. (2011) Environmental barcoding: a next-generation
757 sequencing approach for biomonitoring applications using river benthos. *PLoS One*, **6**, e17497.

758 Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of
759 unsupervised Bayesian clustering. *Bioinformatics*, **27**, 611–618.

760 Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome c oxidase
761 subunit 1 divergences among closely related species. *Proceedings of the Royal Society B:*
762 *Biological Sciences*, **270** Suppl, S96-9.

763 Knowlton N (1993) Sibling species in the sea. *Annual Review of Ecology, Evolution and*
764 *Systematics*, **24**, 189–216.

765 Knowlton N, Jackson JBC (2008) Shifting baselines, local impacts, and global change on coral
766 reefs. *PLoS Biology*, **6**, e54.

767 Lejzerowicz F, Esling P, Pillet LL *et al.* (2015) High-throughput sequencing and morphology
768 perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports*, **5**, 13932.

769 Leray M, Knowlton N (2015) DNA barcoding and metabarcoding of standardized samples reveal
770 patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*, **112**,
771 2076–2081.

772 Leray M, Knowlton N (2016) Censusing marine eukaryotic diversity in the twenty-first century.
773 *Philosophical transactions of the Royal Society of London. Series B*, **371**, 20150331.

774 Leray M, Yang JY, Meyer CP *et al.* (2013) A new versatile primer set targeting a short fragment
775 of the mitochondrial COI region for metabarcoding metazoan diversity: application for
776 characterizing coral reef fish gut contents. *Frontiers in Zoology*, **10**, 34.

777 Liu L, Liu M, Wilkinson DM, Chen H, Yu X, Yang J (2017) DNA metabarcoding reveals that
778 200-µm-size-fractionated filtering is unable to discriminate between planktonic microbial and
779 large eukaryotes. *Molecular Ecology Resources*, in press, doi: 10.1111/1755-0998.12652

780 Logares R, Audic S, Bass D *et al.* (2014) Patterns of rare and abundant marine microbial
781 eukaryotes. *Current Biology,* **24**, 813-821.

782 Marques VM, Reis CS, Calvário J *et al.* (1982) Contribuçao para o estudo dos povoamentos
783 bentónicos (substrato rochoso) da costa occidental portuguesa. Zona intertidal. *Oecologia*
784 *Aquatica*, **6**, 119-145.

785 Massana R, Gobet A, Audic S *et al.* (2015) Marine protest diversity in European coastal waters
786 and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, **17**,
787 4035-4049.

788 McMurdie PJ, Holmes S (2013) phyloseq: an R package for reproducible interactive analysis and
789 graphics of microbiome census data. *PLoS One*, **8**, e61217.

790 Mikkelsen PM, Cracraft J (2001) Marine biodiversity and the need for systematic inventories.
791 *Bulletin of Marine Science*, **69**, 525–534.

792 Munar J (1993) Els equinoderms. In: Alcover JA, Ballesteros E, Fornós JJ, editors. *Història*
793 *Natural de l'Archipèlag de Cabrera*. Palma de Mallorca, Spain. Editorial Moll-CSIC, 597-606.

794 O'Donnell JL, Kelly RP, Lowell NC, Port JA (2016) Indexed PCR primers induce template-
795 specific bias in large-scale DNA sequencing studies. *PLoS One*, **11**, e0148698.

796 Oksanen J, Blanchet FG, Friendly M *et al.* (2016) vegan: Community Ecology Package. R
797 package version 2.4-3. http://CRAN.R-project.org/package=vegan

798 Pawlowski J, Esling P, Lejzerowicz F et al. (2014) Environmental monitoring through protist
799 next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic
800 foraminifera communities. *Molecular Ecology Resources*,**14**, 1129–40.

801 Pearman JK, Anlauf H, Irigoien X, Carvalho S (2016) Please mind the gap – visual census and
802 cryptic biodiversity assessment at central Red Sea coral reefs. *Marine Environmental Research*,
803 **118**, 20–30.

804 Peña V, Bárbara I (2008) Maërl community in the north-western Iberian Peninsula: a review of
805 floristic studies and long-term changes. *Aquatic Conservation: Marine and Freshwater*
806 *Ecosystems*, **18**, 339-366.

807 Ransome E, Geller JB, Timmers M et al. (2017) The importance of standardization for
808 biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS)
809 and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. *PLoS*
810 *One*, **12**, e0175066

811 Ratnasingham S, Hebert PDN (2007) bold: The Barcode of Life Data System
812 (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

813 Reaka-Kudla ML (1997) The global biodiversity of coral reefs: a comparison with rain forests.
814 In: Reaka-Kudla ML, Wilson DE, Wilson EO, editors. *Biodiversity 2: understanding and*
815 *protecting our biological resources,* Joseph Henry Press, 83–108.

816 Rex MA, Ettter RJ (2010) Deep-sea biodiversity. Pattern and scale. Cambridge, Massachusetts.
817 Harvard University Press.

818 Rex MA, Etter RJ, Morris JS *et al.* (2006) Global bathymetric patterns of standing stock and
819 body size in the deep-sea benthos. *Marine Ecology Progress Series*, **317**, 1-8.

820 Sant N, Chappuis E, Rodríguez-Prieto C, Real M, Ballesteros E (2017) Cost-benefit of three
821 different methods for studying Mediterranean rocky benthic assemblages. *Scientia Marina*, **81**,
822 129-138.

823 Taberlet P, Coissac E, Pompanon F *et al.* (2012) Towards next-generation biodiversity
824 assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–50.

825 Tang CQ, Leasi F, Obertegger U *et al.* (2012) The widely used small subunit 18S rDNA
826 molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna.
827 *Proceedings of the National Academy of Sciences,* **109**, 16208–16212.

828 de Vargas C, Audic S, Henry N *et al.* (2015) Eukaryotic plankton diversity in the sunlit ocean.
829 *Science,* **348**, 1261605.

Turon X (1993) Els ascidis: faunística i distribució. In: Alcover JA, Ballesteros E, Fornós JJ, editors. *Història Natural de l'Archipèlag de Cabrera*. Palma de Mallorca, Spain. Editorial Moll-CSIC, 607-621.

Uriz MJ (1993) Les esponges litorals. In: Alcover JA, Ballesteros E, Fornós JJ, editors. *Història Natural de l'Archipèlag de Cabrera*. Palma de Mallorca, Spain. Editorial Moll-CSIC, 531-547.

Wangensteen OS, Cebrian E, Palacín C, Turon X (in press) Under the canopy: Community-wide effects of invasive algae in Marine Protected Areas revealed by metabarcoding. *Marine Pollution Bulletin*.

Wangensteen OS, Turon X (2017) Metabarcoding techniques for assessing biodiversity of marine animal forests. In: Rossi S, Bramanti L, Gori A, Orejas C, editors. *Marine animal forests. The ecology of benthic biodiversity hotspots,* Springer International Publishing, Switzerland, pp. 445-473, ISBN: 978-3-319-21011-7

Warwick RM, Joint IR (1987) The size distribution of organisms in the Celtic Sea: from bacteria to Metazoa. *Oecologia*, **73**, 185-191.

Wheeler QD, Raven PH, Wilson EO (2004) Taxonomy: impediment or expedient? *Science,* **303**, 285.

Zabala M (1993) Els briozous. In: Alcover JA, Ballesteros E, Fornós JJ, editors. *Història Natural de l'Archipèlag de Cabrera*. Palma de Mallorca, Spain. Editorial Moll-CSIC, 561-577.

Data Accessibility

The resulting datasets for COI and 18S, including sequences, taxonomic assignment and abundances for all MOTUs in every sample, have been deposited in the Mendeley data repository (https://data.mendeley.com/datasets/nm2c97fjng/1). Databases of reference sequences to be used for ecotag taxonomic assignment of COI and 18S are deposited in Github (http://github.com/metabarpark/Reference-databases). R scripts used as part of the analysis pipeline are also deposited in Github (http://github.com/metabarpark/R_scripts_metabarpark).

Table 1(on next page)

MOTU richness and morphospecies richness.

Comparison of MOTU richness values obtained in the present work with morphospecies diversity found with morphological methods in previous studies on the same or similar and geographically close communities.

1

Studied area	Community	Group	Morphospecies richness	MOTU richness 18S	MOTU richness COI	References
Ria de Vigo	Cíes detritic	Rhodophyta Phaeophyceae Chlorophyta	123 19 18	303 35 15	842 156 5	Peña & Bárbara 2008
NW Portugal	Cíes *C. tamariscifolia*	Rhodophyta Phaeophyceae Chlorophyta Polychaeta Crustacea Mollusca Echinodermata	30 12 4 45 35 9 4	183 50 10 210 270 143 23	470 178 4 258 879 604 17	Marques et al. 1982
Balearic Islands	Cabrera precoralligen	Rhodophyta Phaeophyceae Chlorophyta	85 12 5	235 41 11	747 107 3	Ballesteros 1994
Mediterranean Spain	Cabrera detritic	Rhodophyta Phaeophyceae Chlorophyta	197 44 31	268 34 12	623 88 4	Peña & Bárbara 2008
Cabrera archipelago	Cabrera all communities	Macroalgae Porifera Cnidaria Bryozoa Decapoda Mollusca Ascidiacea Echinodermata	277 98 85 145 69 169 41 53	458 250 310 290 47 138 50 25	1676 726 1336 696 32 375 138 28	Ballesteros 1993 Uriz 1993 Gili et al. 1993 Zabala 1993 Corbera et al. 1993 Altaba 1993 Turon 1993 Munar 1993

2

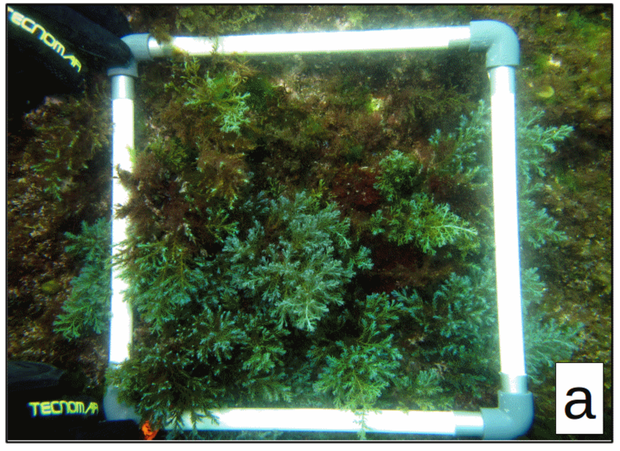
3

4

Figure 1(on next page)

Natural benthic communities sampled in this study.

(a) photophilous community with *Cystoseira tamariscifolia*, (b) photophilous community with *Cystoseira nodicaulis*, (c) sciaphilous with *Saccorhiza polyschides*, (d) Atlantic detritic bottoms (maërl), (e) photophilous community with the invasive seaweed *Lophocladia lallemandii*, (f) photophilous Mediterranean seaweeds, (g) sciaphilous precoralligenous outcrops, (h) Mediterranean detritic bottoms (maërl). (a)-(d) from Galician Atlantic Islands National Park, (e)-(f) from Cabrera Archipelago National Park. In (a) the 25-cm quadrat sampling unit is shown.



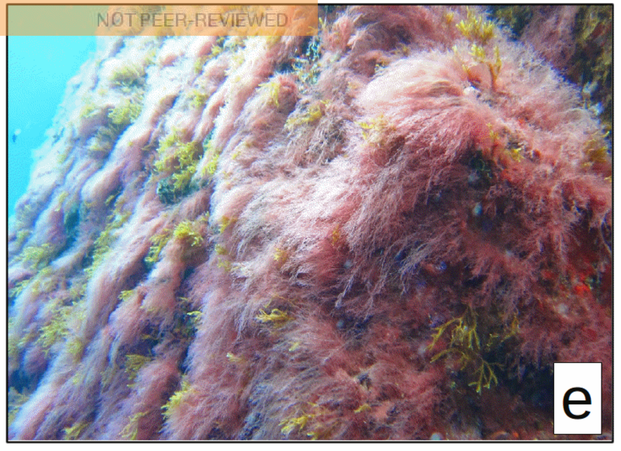


Figure 2

Percentages of *in silico* taxonomic coverage per species using the new Leray-XT primer set for COI.

Figures are compared to the original Leray primer set (Leray et al. 2013), for the main eukaryotic groups (above) and the metazoan phyla (below).

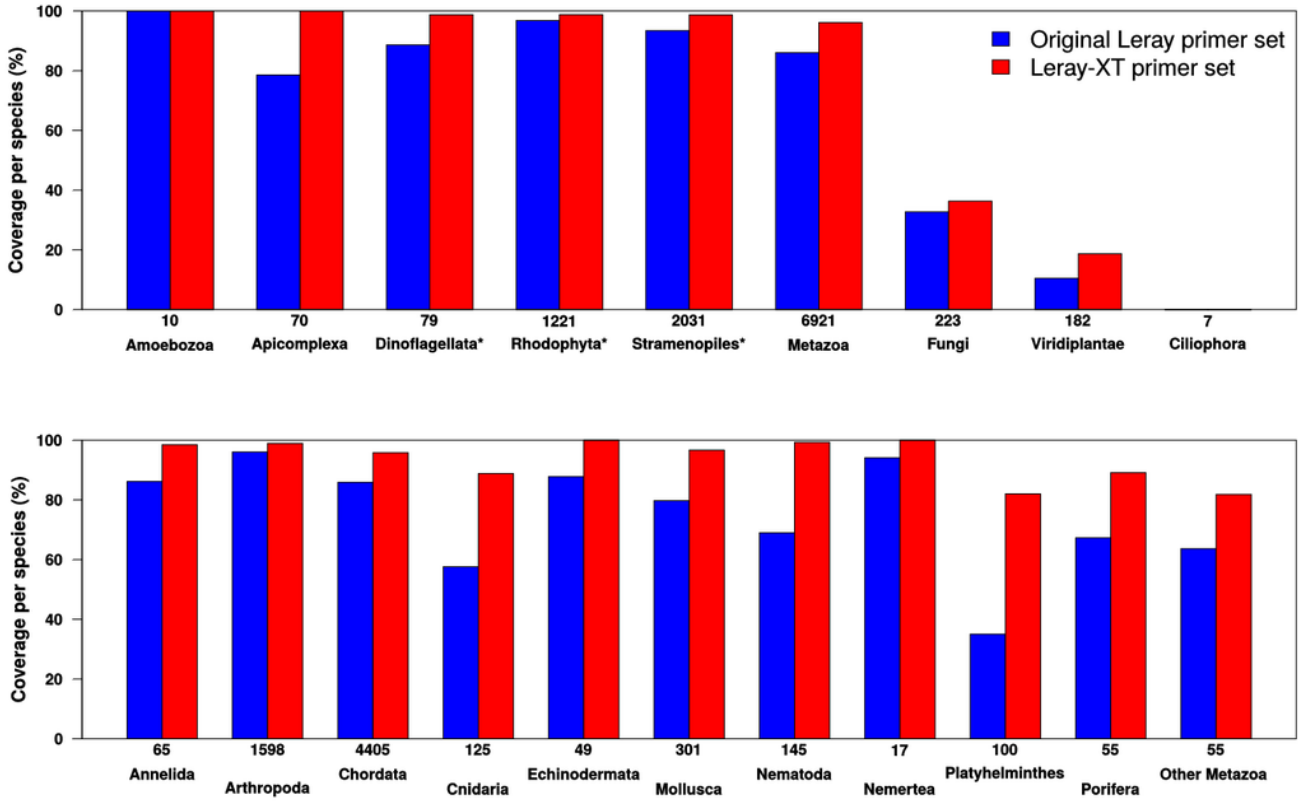
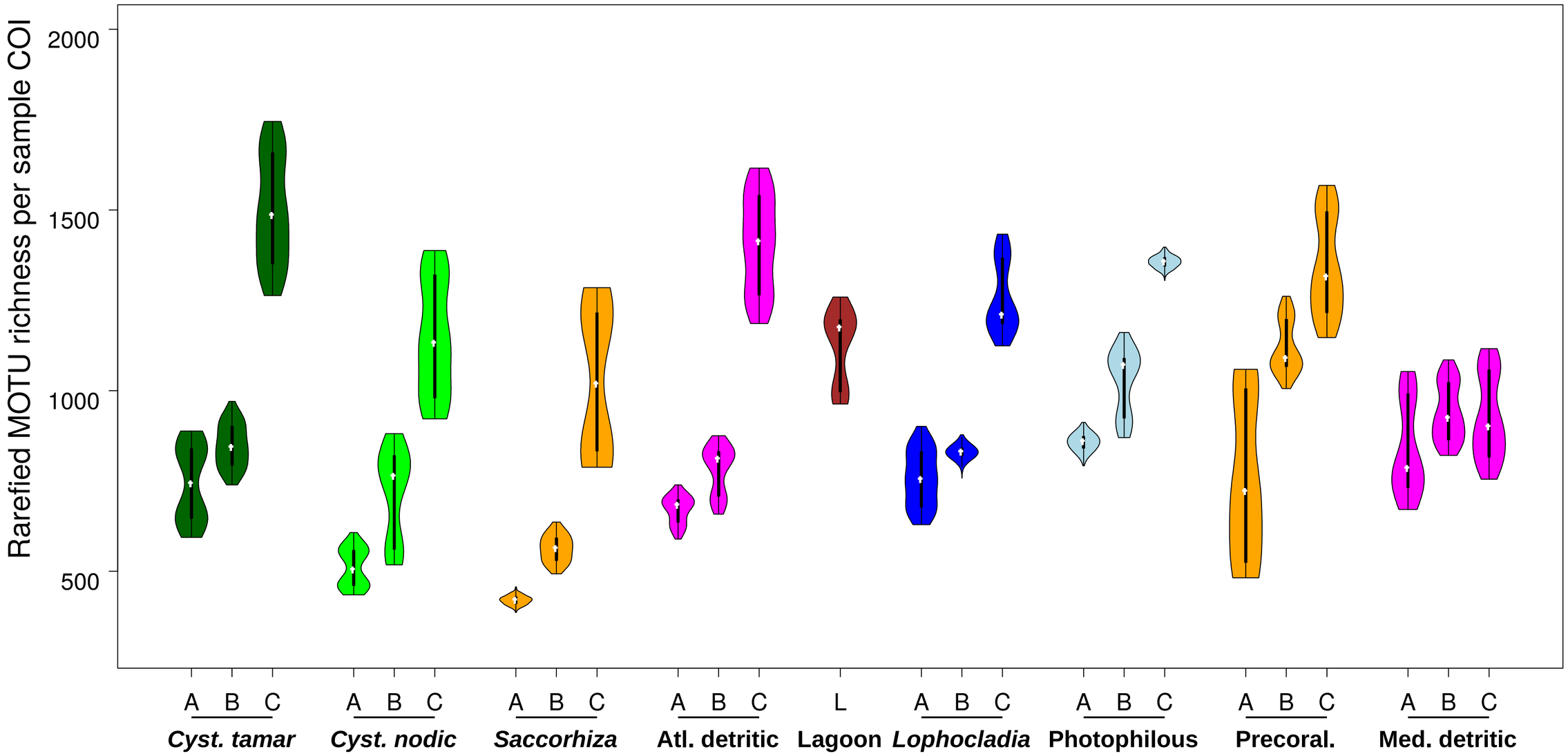


Figure 3(on next page)

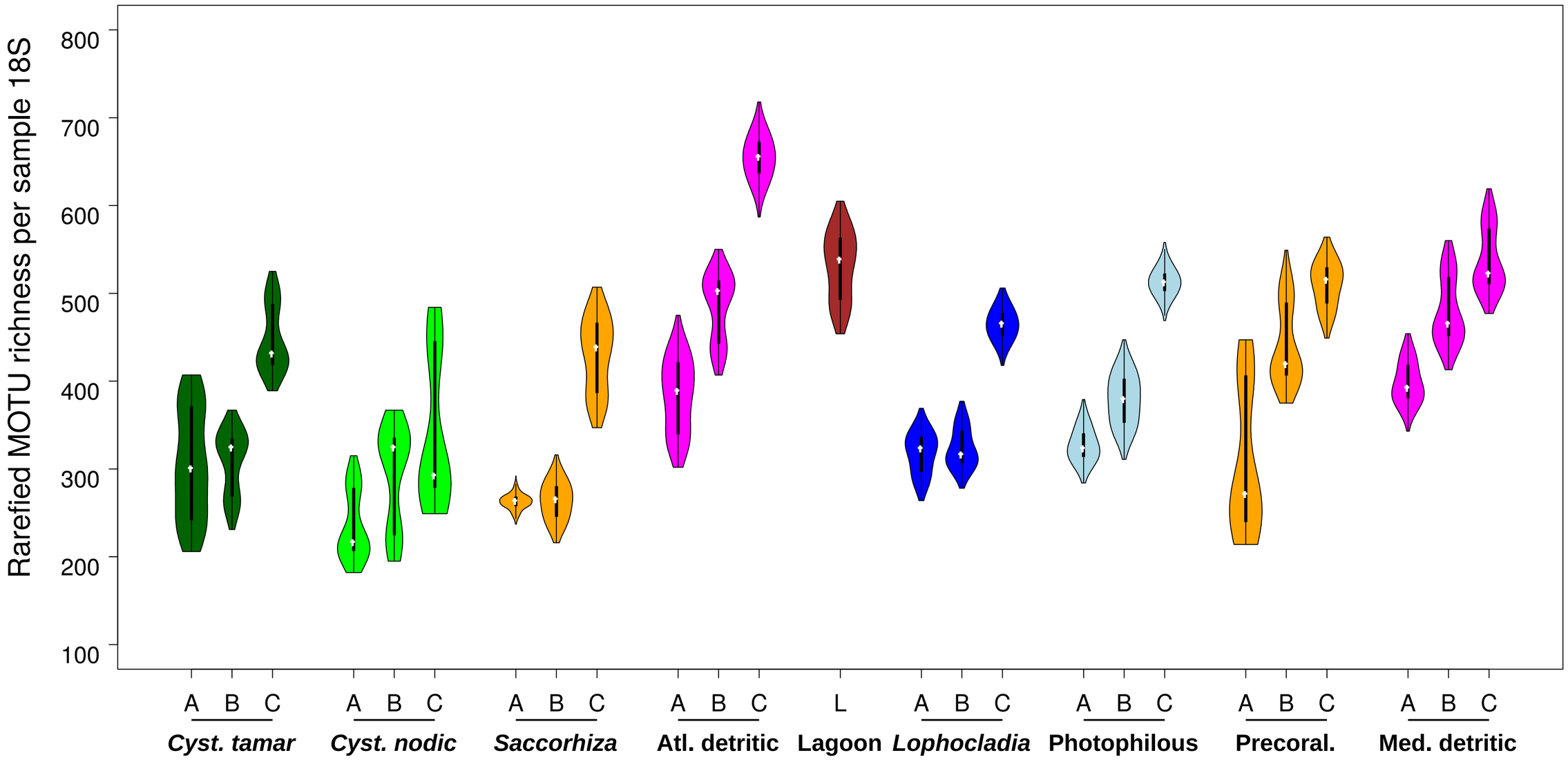
Violin plots showing patterns of α-diversity.

Graphs of three different size fractions (A, B and C) in eight different hard substrate marine communities and a set of tidal lagoon sediment samples for comparison purposes. Results obtained by rarefaction analysis to 10,000 reads per sample with 500 replicates. a: MOTU richness obtained from COI. b: MOTU richness obtained from 18S.

a



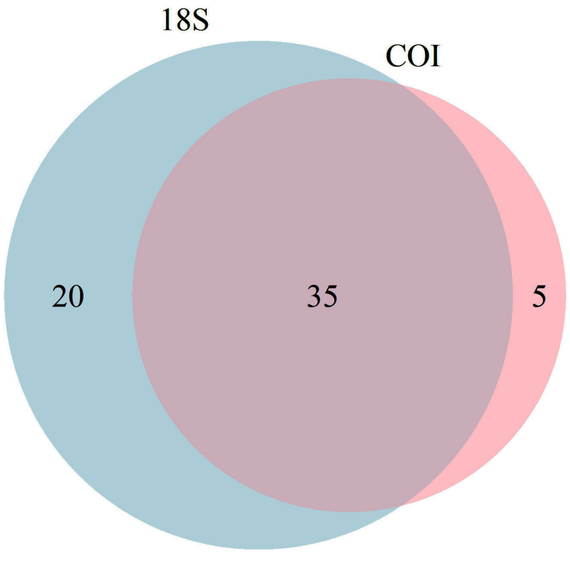
b



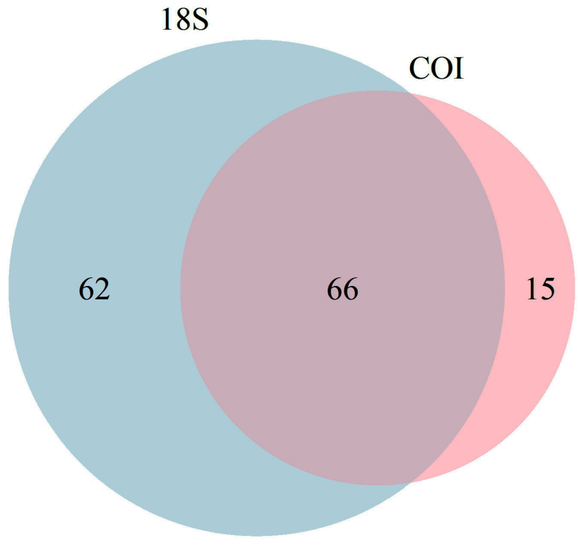
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3429v1 | CC BY 4.0 Open Access | rec: 23 Nov 2017, publ: 23 Nov 2017

Figure 4(on next page)

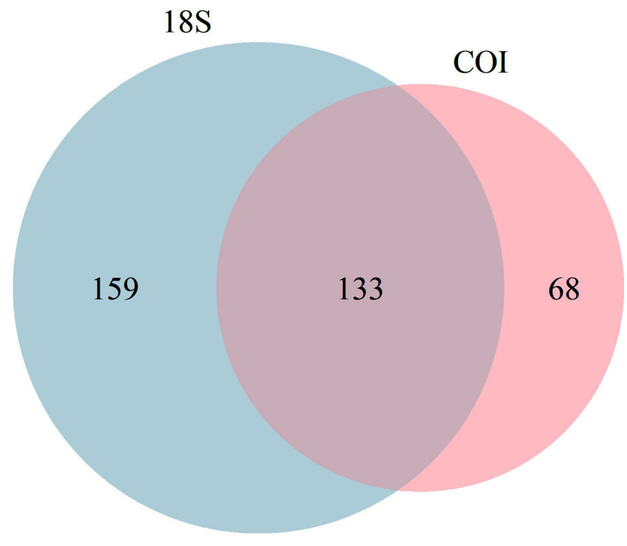
Venn diagrams comparing the number of different taxa recovered using COI (red) or 18S (blue) from all studied communities, for different taxonomic ranks .



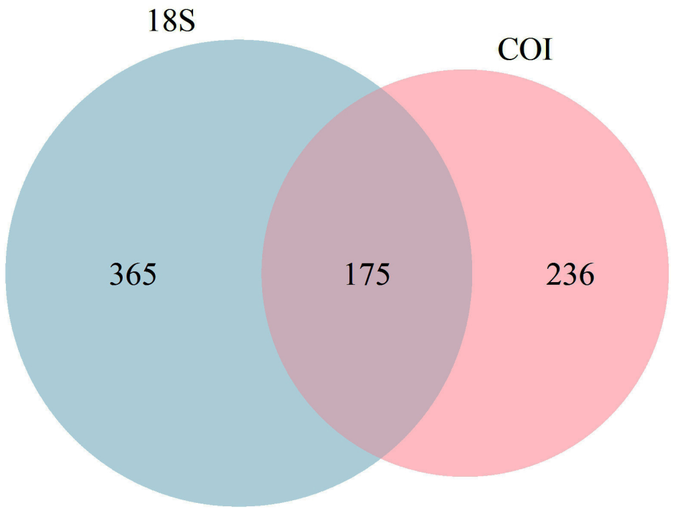
Phylum



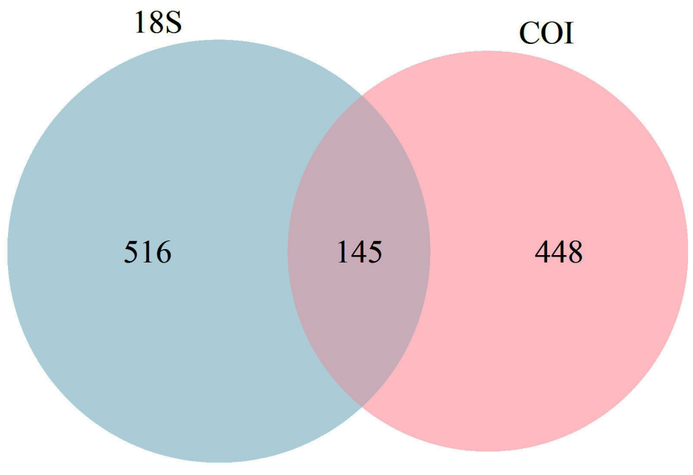
Class



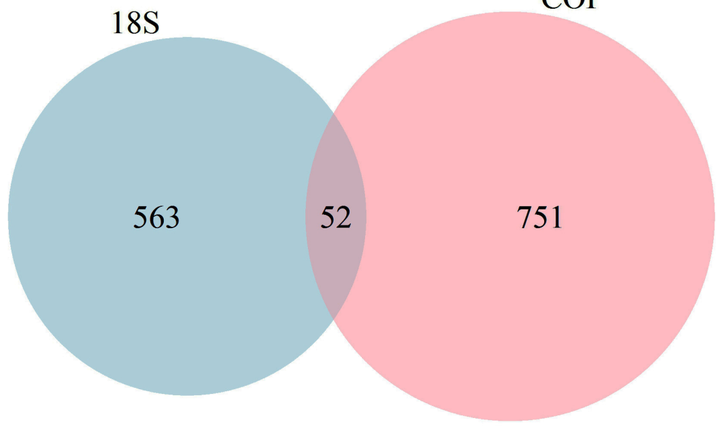
Order



Family



Genus



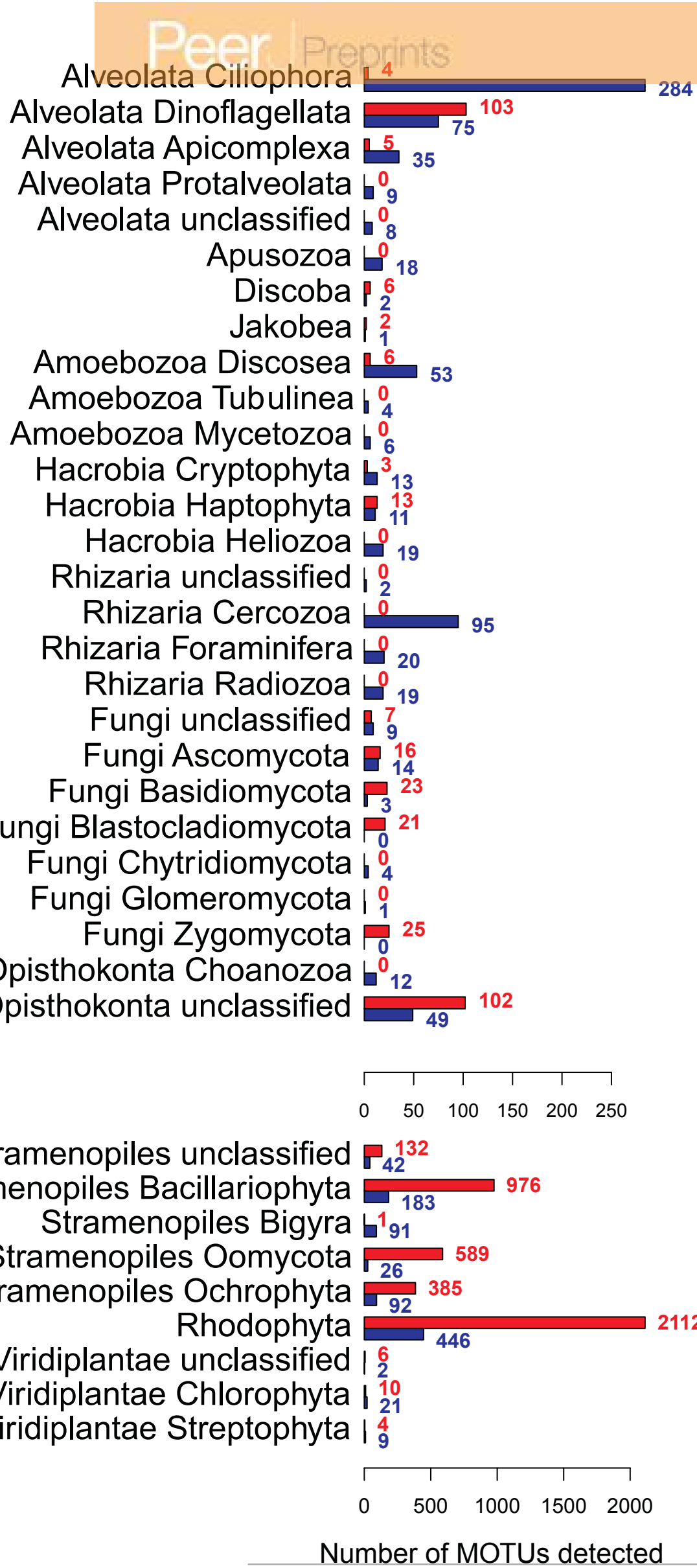
species

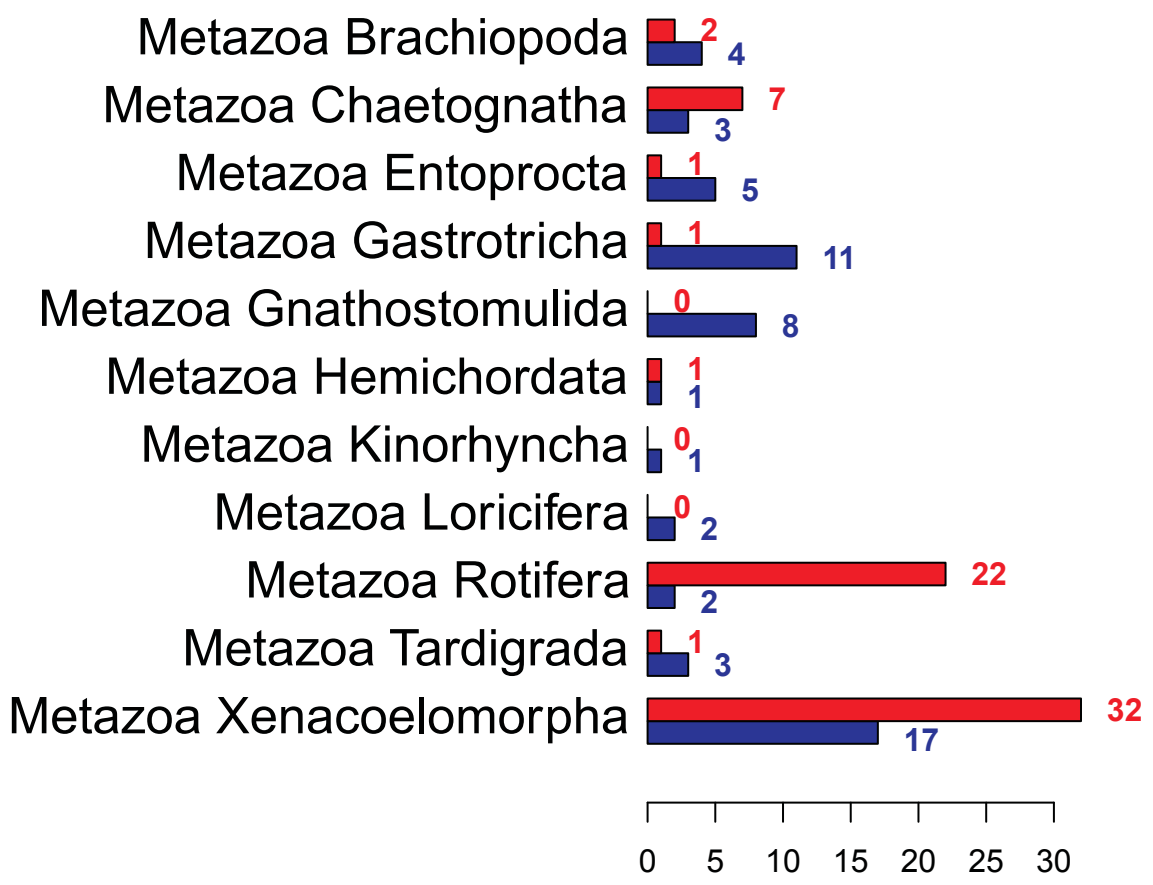
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3429v1 | CC BY 4.0 Open Access | rec: 23 Nov 2017, publ: 23 Nov 2017

Figure 5(on next page)

Number of MOTUs detected for every phylum in the studied communities using COI (red) or 18S (blue) metabarcoding.

Note the different scales used for small-sized groups (protists, fungi and small metazoans) and for dominant groups in benthic communities (metazoans and seaweeds).





PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3429v1 | CC BY 4.0 Open Access | rec: 23 Nov 2017, publ: 23 Nov 2017

Figure 6

Patterns of abundance of metabarcoding reads per community and fraction size.

Results obtained using COI (a) or 18S (b) in eight different marine littoral communities from NE Atlantic (left) and W Mediterranean (right), and a set of lagoon sediment samples. All replicates from the same community and fraction size have been pooled. See supplementary fig. S4 for the same figure, split by replicates.

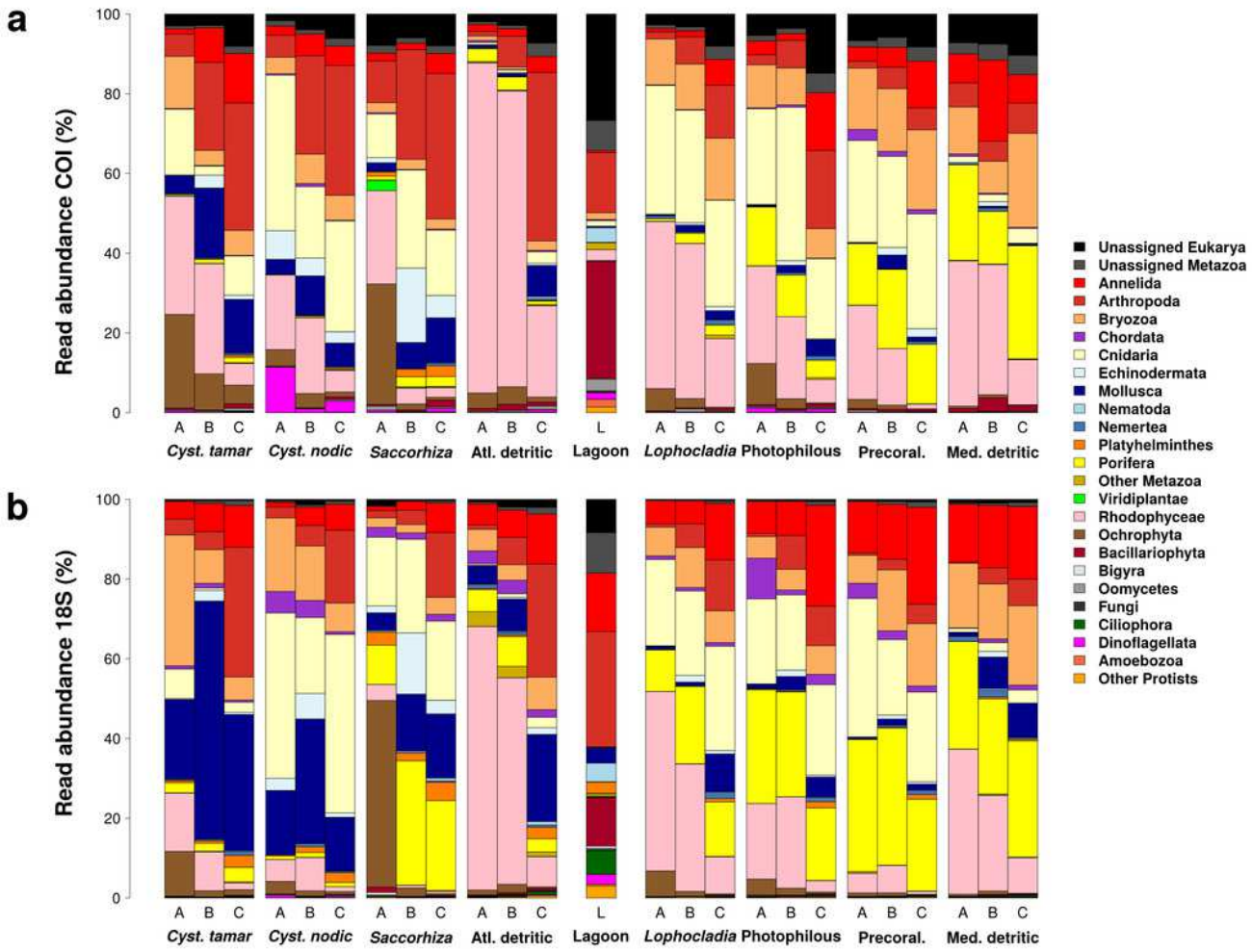


Figure 7

Patterns of relative MOTU richness using COI (a) or 18S (b) in the same communities of fig.6. See supplementary fig. S5 for the same figure, split by replicates.

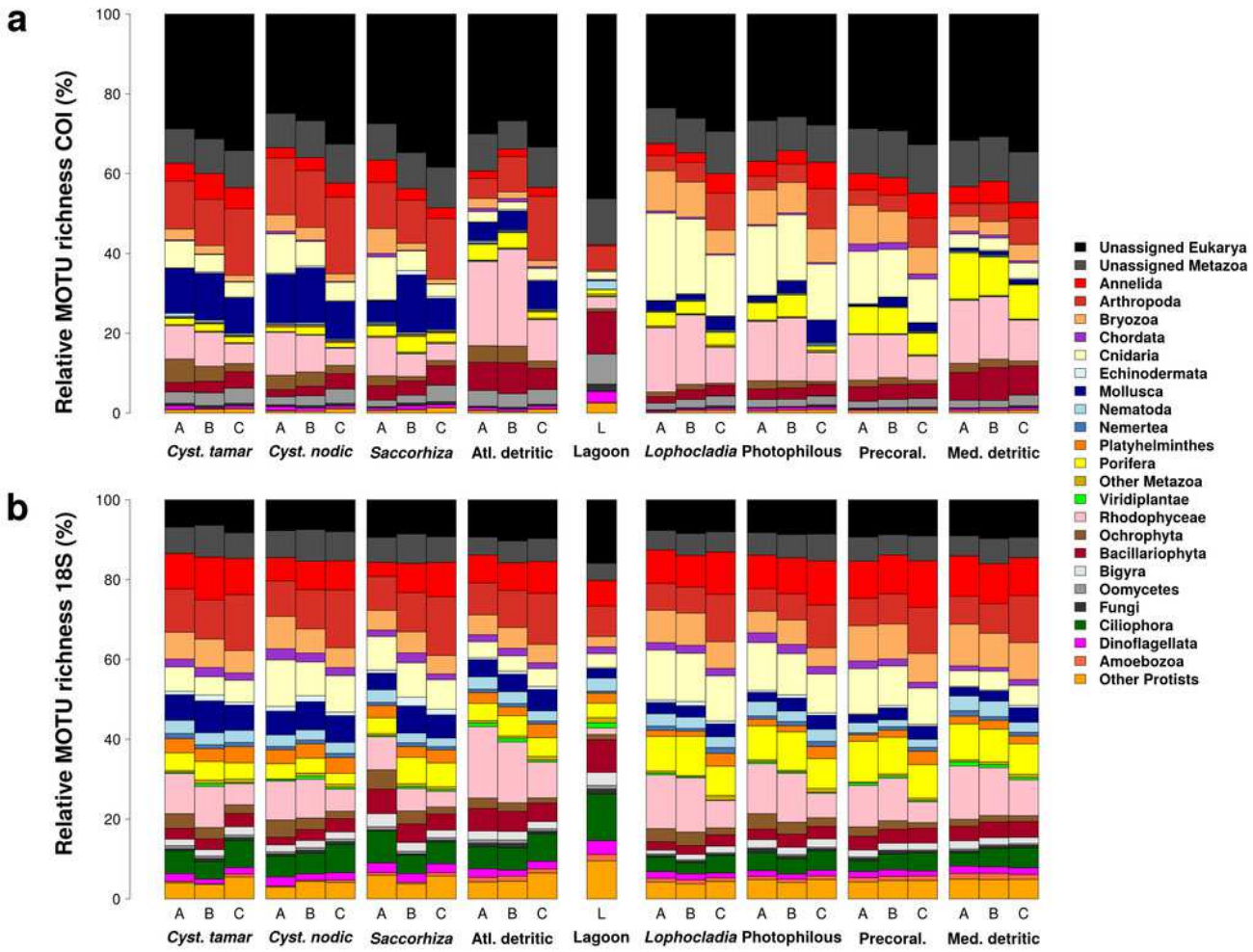
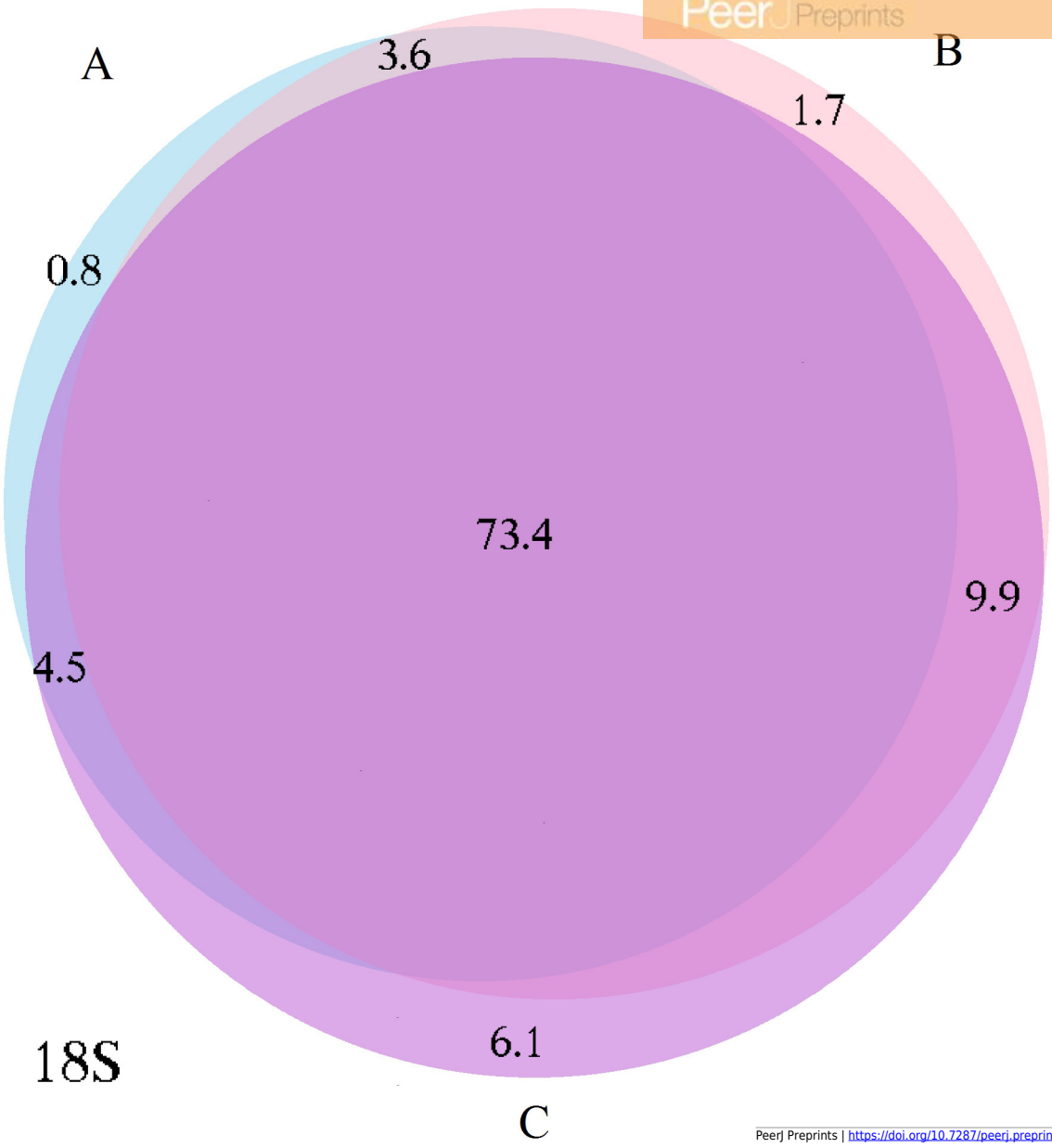
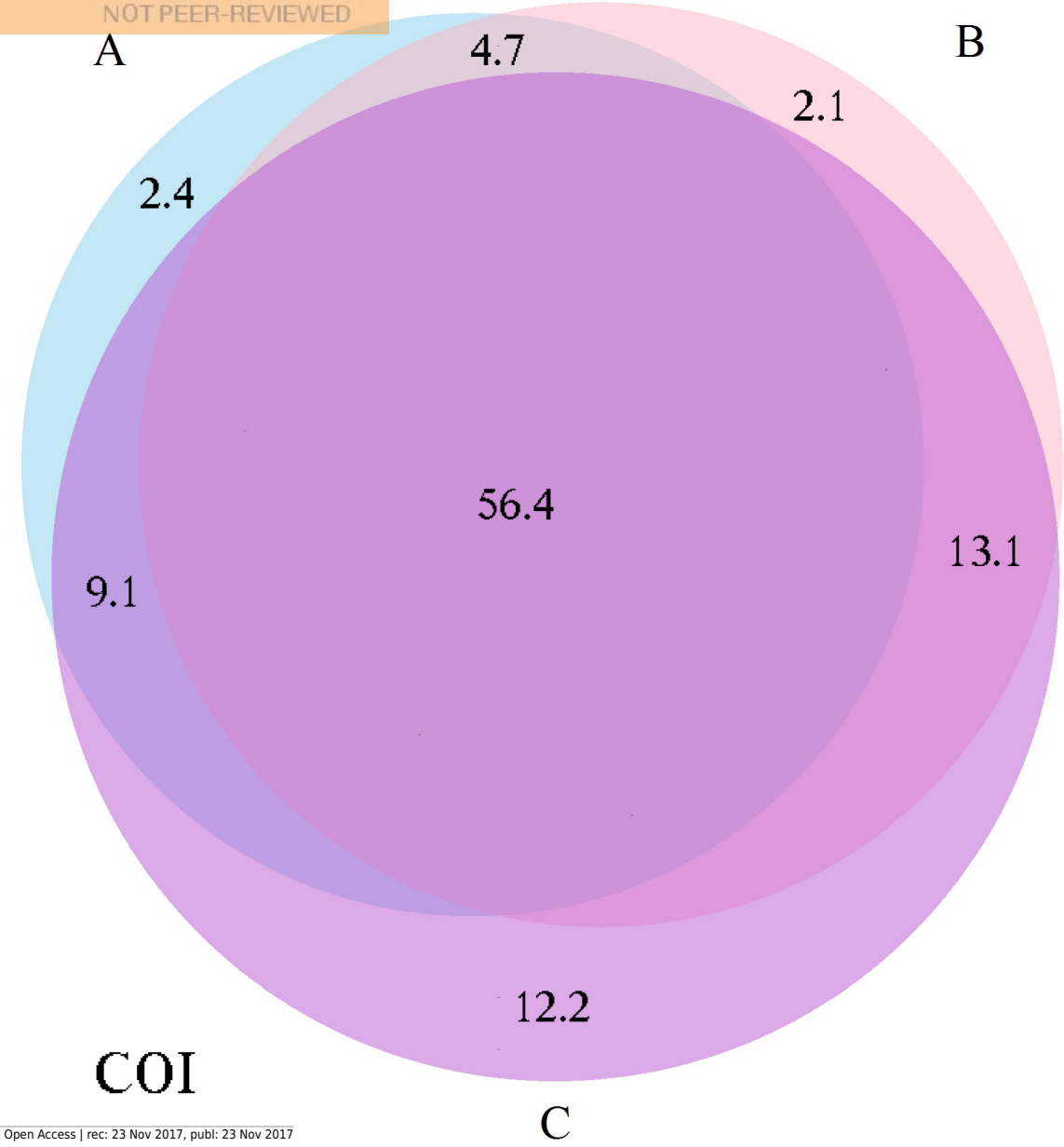


Figure 8(on next page)

Venn diagram showing the number of MOTUs detected in each size fraction (A, B or C) for 18S (left) and COI (right) in all communities. Numbers are percentages of total MOTUs.



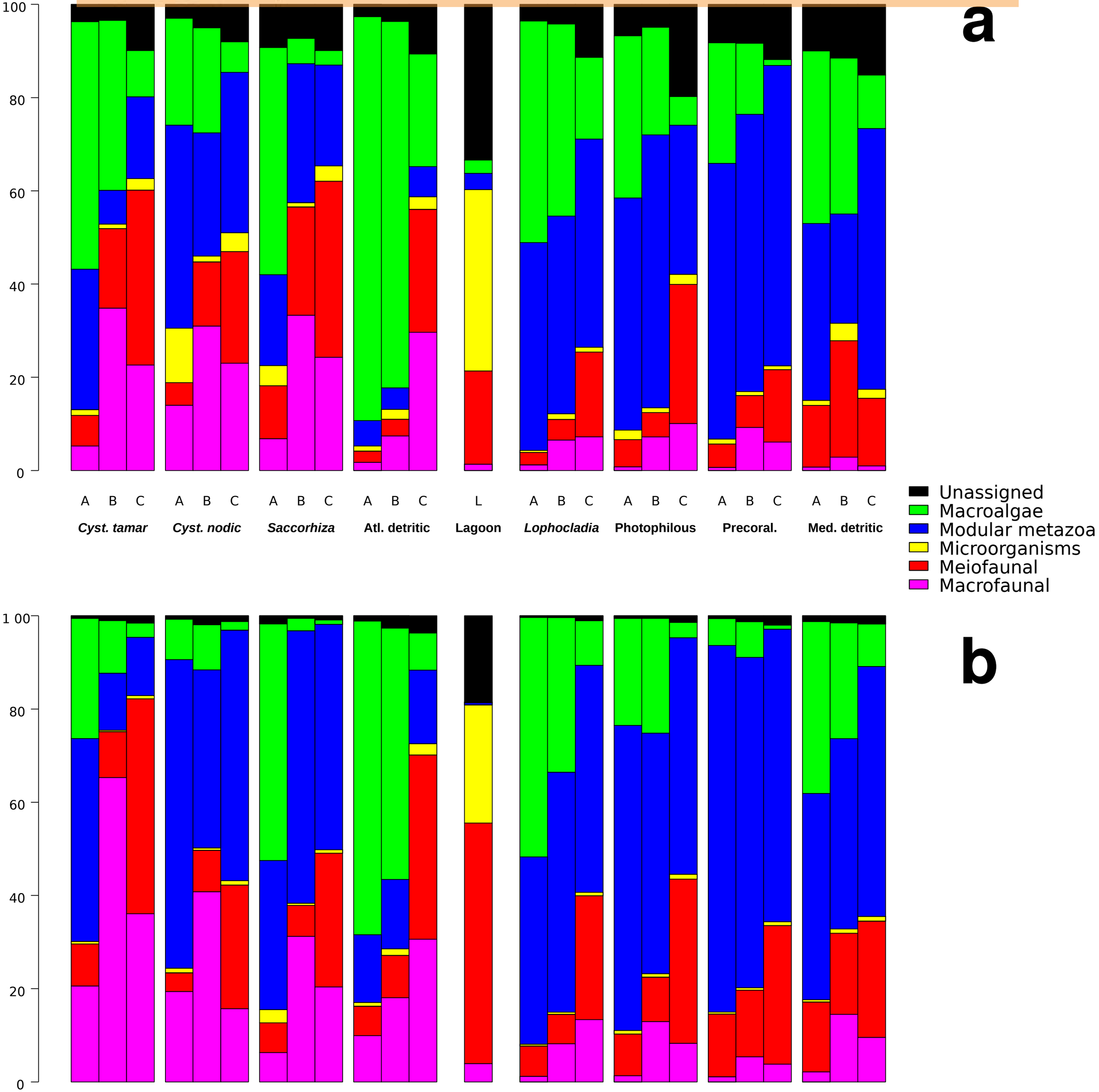


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3429v1 | CC BY 4.0 Open Access | rec: 23 Nov 2017, publ: 23 Nov 2017

Figure 9(on next page)

Effect of size fractionation in the recovery of different ecological categories by metabarcoding.

Results using COI (a) and 18S (b) on eight different littoral communities from national parks and sediment samples from a tidal lagoon.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3429v1 | CC BY 4.0 Open Access | rec: 23 Nov 2017, publ: 23 Nov 2017